METHODS FOR FUNCTIONAL INFERENCE IN THE PROTEOME AND INTERACTOME

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Michael Joseph Meyer

January 2017

METHODS FOR FUNCTIONAL INFERENCE

IN THE PROTEOME AND INTERACTOME

Michael Meyer

Cornell University 2017

Over the past several decades, biology has become an increasingly data-driven science. Due in large part to new techniques that allow massive collection of biological data, including next-generation sequencing and high-throughput experimental screening, many of the limitations currently facing the field are in the organization and interpretation of these data. In this dissertation, I present several computational methods and resources designed to organize and perform functional inference on these systems-level biological data sources. In Chapters 2 and 3, I describe the construction of a database and web tool to aid in foundational genomics research by providing predictions of interacting protein domains in interactomes and all-by-all conversions of popular variant identification formats. In Chapter 4, I describe the construction of the first whole-interactome protein interaction network in the fission yeast *S. pombe*, and, through comparisons with other complete networks in human and the budding yeast S. cerevisiae, demonstrate principles of functional evolution. Finally, in Chapters 5 and 6, I propose two new methods for functional genomic inference—an algorithm to predict cancer driver genes and mutations through 3D atomic clustering of somatic mutations and an ensemble machine learning method to predict the 3D interfaces of protein interactions by taking into account the evolutionary relationships and biophysical properties of proteins. Taken together, this suite of computational resources will help researchers interpret biological function on a genomic scale.

BIOGRAPHICAL SKETCH

Michael was born in 1989 to Joseph and Nicolette Meyer, and grew up in Worthington, Ohio as the eldest of three children. He graduated from Thomas Worthington High School in 2007 and attended the Massachusetts Institute of Technology from 2007 to 2011, where he earned his bachelor's degree in Biological Engineering. In the summer of 2010 he interned at Ion Torrent, a startup DNA sequencing company founded by Dr. Jonathan Rothberg, where he helped develop the Ion Torrent Personal Genome Machine. He returned in the summer of 2011 after a successful buyout by Life Technologies.

In 2011 he married his high school sweetheart, Sarai Meyer (Itagaki), and they moved to Ithaca, New York to begin their Ph.D.s at Cornell University. Michael entered the Tri-Institutional Training Program in Computational Biology and Medicine (CBM) and completed laboratory rotations at Cornell's Ithaca campus and at the Weill Cornell Medical School in New York City. He ultimately joined the computational systems biology lab of Dr. Haiyuan Yu in the Weill Institute for Cell and Molecular Biology in Ithaca, where he completed the research presented in this dissertation.

For my grandfathers

ACKNOWLEDGEMENTS

I am incredibly fortunate to have had the support of many people over the course of my life and studies. There are an awful lot of people to thank. But first, I'd like to acknowledge my incredible fortune to have been born in a time and place where scholarly aspirations are relevant and valued, without which none of this would have been possible.

To my parents—Mom and Dad, thank you for giving me every chance to succeed, for investing in me, and for showing me how to be a good person. Mom, thank you for your unconditional love, home-cooked meals, and for giving me that problem-solving itch that I'm still trying to scratch. Dad, thank you pushing me, allowing me to learn from my mistakes, accept failure, and think for myself. And thank you both for being there from a distance to help me through these past five years. And to the rest of my family—thanks for putting up with me, for caring about my well-being, and for occasionally mustering the bravery to ask about my research.

To all of my teachers throughout the years—Kathy McCulloch, Mark Hill, Peter Scully, and so many others, thank you for inspiring me. To my most recent mentor, Haiyuan, thank you for entrusting me with a great deal of responsibility in the early years of your lab, and for providing advice when I needed it, but still allowing me the freedom to determine my own course of study and explore my passions. To my committee—Chris, David, and Olivier—thank you for you for your genuine interest in my work and your feedback over the years.

To several generations of labmates—thank you for being my daily buffer against the slog of graduate school—for the laughs, distractions, and taunts during March Madness. To the amazing undergraduate researchers I've worked with—Ryan, Philip, Mark, and Aaron—thank you for your enthusiasm, willingness to learn, and major contributions to much of the

work in this dissertation. To the old guard—Jishnu, Amanda, Tommy, and Robert—thanks for being there from the beginning and for helping to establish a productive and supportive research environment. Jishnu, thanks for showing me the ropes, for the most interesting arguments, and for sharing your encyclopedic knowledge of the literature. Tommy, thanks for your tireless efforts to make FissionNet a success, and for teaching me a thing or two about the wet lab in the process. To the new kids—Juan, Siwei, Antoine, and Charles—good luck. Juan, thank you for arriving at just the right time to remind me why I love what I do, for all the white-boarding, and for helping to create some really cool stuff. Thanks for the late-night code jams and after-work beers, for provocative discussions of philosophy, and for one awkward bro hug.

And to Sarai—thanks for, well, everything. For all the little stuff—slipping butter between two halves of a grab-and-go muffin on a Monday morning, counting cats on summer evening walks, and leaving me little notes to find around the house. Thank you for being there, for smiling and laughing, for spontaneity and routine. And thank you for going with me on this first of many adventures—we made it.

CONTRIBUTIONS

I would also like to acknowledge the work of others that has gone into each of the chapters presented in this dissertation. Due to the collaborative nature of biological research, I have had the pleasure to work with many gifted scientists, several of whom are co-authors on the publications presented here. (* indicates co-first author)

Chapter 2: Meyer, M.J.*, Das, J.*, Wang, X.*, and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29, 1577-1579.

Chapter 3: Meyer, M.J.*, Geske, P.*, and Yu, H. (2016). BISQUE: locus- and variant-specific conversion of genomic, transcriptomic and proteomic database identifiers. *Bioinformatics* 32, 1598-1600.

Chapter 4: Vo, T.V.*, Das, J.*, Meyer, M.J.*, Cordero, N.A., Akturk, N., Wei, X., Fair, B.J., Degatano, A.G., Fragoza, R., Liu, L.G., et al. (2016). A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. *Cell* 164, 310-323.

Chapter 5: Meyer, M.J.*, Lapcevic, R.*, Romero, A.E.*, Yoon, M., Das, J., Beltran, J.F., Mort, M., Stenson, P.D., Cooper, D.N., Paccanaro, A., and Yu H. (2016). mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Human Mutation* 37, 447-456.

Chapter 6: Meyer, M.J.*, Beltran, J.F.*, Fragoza, R., Rumack, A., and Yu, H. (2016). A pan-interactome map of protein interaction interfaces. *Under Review*.

TABLE OF CONTENTS

Appendices

CHAPTER 1

Introduction: data and inference in the genomics era

1.1 A biological data explosion

Since the completion of the human genome project (Lander et al., 2001), the precipitous drop in the cost of DNA sequencing has allowed the wholesale sequencing of new genomes and placed targeted sequencing for small scale studies within the reach of even modestly sized laboratories. This has brought about an explosion of sequencing data in the form of genomes of new organisms, human population variation studies, and clinical sequencing screens. While it took nearly 15 years to completely sequence one human genome, today there are ongoing projects to sequence the full exomes of entire populations, with over 60,000 humans already sequenced (Lek et al., 2016). In the clinic, sequencing is now used regularly to investigate underlying genetic aberrations associated with cancers, and deposition of whole genome cohort studies as well as small scale studies has led to the discovery of vastly over 1 million somatic mutations associated with cancer (Forbes et al., 2015).

Experimental technologies have also added to our knowledge of protein interactions and protein structures through a combination of advancements in robotics that make possible genomic-scale, high-throughput screens and a steady accumulation of experimentally determined structures in public repositories. For instance, just 15 years ago there were no fully screened interactomes available for any organism. Today, there are three organisms with complete or nearly complete interactomes (Rolland et al., 2014; Yu et al., 2008), including the recently screened *S. pombe* interactome described in Chapter 4 (Vo et al., 2016), and several other interactomes have been partially screened (Arabidopsis Interactome Mapping Consortium, 2011; Giot et al., 2003; Stelzl et

al., 2005). We also now know many more structures of proteins from X-ray crystallography, NMR, and more recently cryo-EM (Kuhlbrandt, 2014)—today the Protein Databank contains over 115,000 experimentally determined structures of proteins compared to only ~10,000 just 16 years ago (Berman, 2000).

These new technologies, including increased efficiency of some classic biological research techniques, have been a great boon for traditional hypothesis-driven research. However there are also new opportunities for genomic-scale systems biology research to begin to detect functional patterns that can help us understand the workings of biological systems.

1.2 New challenges

The scientific challenges now facing biology demand new methods for organizing, storing, disseminating, and analyzing this deluge of data. The past decade has seen the emergence of many computational resources to manage these data, especially online resources such as the UCSC Genome Browser (Karolchik et al., 2014), UniProt (UniProt-Consortium, 2015), PDB (Berman, 2000), Ensembl (Cunningham et al., 2015), COSMIC (Forbes et al., 2015), and many others. Still, in order to perform research based on these data, new methods are needed to perform analysis and make predictions of biological function. To aid in this effort, many resources that make available these analyses and predictions are also available, including predictions of protein binding interfaces (Meyer et al., 2013; Mosca et al., 2013) and of 3D protein structures (Pieper et al., 2011).

Broad strategies for identifying and interpreting trends in biological data have also emerged as scientists have embraced techniques from statistics and computer science. For instance, statistical and machine learning techniques are now pervasive in biology, and are used for predicting the deleteriousness of mutations (Adzhubei et al., 2010), cancer driver genes (Lawrence et al., 2013)

and prognosis (Das et al., 2015), and protein binding interfaces (Meyer et al., 2013; Mosca et al., 2013), among many others. In this work, I present several such methods and accompanying web services—in Chapters 2 and 6, I present methods for predicting binding interfaces of protein interactions (Meyer et al., 2013; Meyer et al., 2016a), and, in Chapter 5, I present a method to predict driver mutations and genes in cancer (Meyer et al., 2016c).

1.3 Perspectives for the future

The democratization of data through online resources has been a key step in ushering in the genomic era; however, many challenges still need to be met to ensure that meaningful scientific discoveries can be made. Unfortunately, the rapid pace of development of many computational resources sometimes fosters complacency of both developers and users, leaving behind tools that in the best case may be difficult to use and interpret, and in the worst case become completely irrelevant in a short period of time.

One key challenge is maintaining compatibility of databases storing the same types of data using different naming conventions. This is especially problematic for sequencing data, as there can be disagreement over names and sequence of genes, transcripts, and proteins. Ideally, there will be future consolidation of databases and conventions, leading to more reliable sharing of data. In the meantime, methods to convert between different naming conventions are needed to allow collaboration and oversight. In Chapter 3, I present a web tool to provide conversion of variant representations, which, due to their annotation in relation to genes, transcripts, or proteins, provide an extra challenge for researchers (Meyer et al., 2016b).

Another challenge is the need to determine the quality and source of data available in databases. This is especially concerning in an age when a researcher's visit to a web resource may occur before

or even in lieu of reading the associated scientific publication. Without knowledge of accession procedures for databases, low-quality or predicted data may be easily confused with high-confidence experimentally determined data. This is often made even more confusing for users based on shared file formats for data of vastly different confidence—for instance, predicted protein homology models use the same coordinate file structure as experimentally determined structures (Berman, 2000; Pieper et al., 2011). For interactome datasets, this can also be an issue, as using predicted interactions to validate functional hypotheses can be prone to logical circularities—e.g. we cannot determine conservation of interactomes by using interactions predicted using conservation.

However standards are continually being laid out by the community—some leaders in this effort include Ensemble, NCBI, and UniProt—and researchers are showing a new interest in the implications of widely accessible data as a topic of primary research (Altman, 2004; Bourne, 2005; Stein, 2003). Due to the pervasiveness of large biological data, I expect a natural increase in responsible use of these data and understanding of data sources. In the meantime, we can, as with any field of research, look to the giants of the field to provide guidance-by-example for best practices, but we must also take responsibility ourselves for determining the proper use of these resources. The biological knowledge that we stand to gain through analyses of these data makes all of these efforts undoubtedly worthwhile.

1.4 References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Altman, R.B. (2004). Editorial: Building successful biological databases. Briefings in bioinformatics *5*, 4-5.

Arabidopsis Interactome Mapping Consortium (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Berman, H.M. (2000). The Protein Data Bank. Nucleic Acids Research *28*.

Bourne, P. (2005). Will a biological database be different from a biological journal? PLoS computational biology *1*, 179-181.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.*, et al.* (2015). Ensembl 2015. Nucleic Acids Res *43*, D662-669.

Das, J., Gayvert, K.M., Bunea, F., Wegkamp, M.H., and Yu, H. (2015). ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers. BMC genomics *16*, 263.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S.*, et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res *43*, D805-811.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E.*, et al.* (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727-1736.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res *42*, D764-770.

Kuhlbrandt, W. (2014). Cryo-EM enters a new era. eLife *3*, e03678.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lawrence, M., Stojanov, P., Polak, P., Kryukov, G., Cibulskis, K., Sivachenko, A., Carter, S., Stewart, C., Mermel, C., Roberts, S.*, et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Meyer, M., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics (Oxford, England).

Meyer, M.J., Beltran, J.F., Fragoza, R., Rumack, A., and Yu, H. (2016a). A pan-interactome map of protein interaction interfaces. Under Review.

Meyer, M.J., Geske, P., and Yu, H. (2016b). BISQUE: locus- and variant-specific conversion of genomic, transcriptomic and proteomic database identifiers. Bioinformatics *32*, 1598-1600.

Meyer, M.J., Lapcevic, R., Romero, A.E., Yoon, M., Das, J., Beltran, J.F., Mort, M., Stenson, P.D., Cooper, D.N., Paccanaro, A.*, et al.* (2016c). mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum Mutat *37*, 447-456.

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nature methods *10*, 47-53.

Pieper, U., Webb, B., Barkan, D., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E., Pettersen, E., Huang, C.*, et al.* (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research *39*, 74.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212-1226.

Stein, L.D. (2003). Integrating biological databases. Nat Rev Genet *4*, 337-345.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

UniProt-Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

Vo, T.V., Das, J., Meyer, M.J., Cordero, N.A., Akturk, N., Wei, X., Fair, B.J., Degatano, A.G., Fragoza, R., Liu, L.G., *et al.* (2016). A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. Cell *164*, 310-323.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., *et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

CHAPTER 2

INstruct: a database of high quality three-dimensional structurally resolved protein interactome networks

2.1 ABSTRACT

INstruct is a database of high-quality, three-dimensional, structurally resolved protein interactome networks in human and six model organisms. INstruct combines the scale of available high-quality binary protein interaction data with the specificity of atomic-resolution structural information derived from co-crystal evidence using a tested interaction interface inference method. Its web interface is designed to allow for flexible search based on standard and organism-specific protein and gene-naming conventions, visualization of protein architecture highlighting interaction interfaces, and viewing and downloading custom three-dimensional structurally resolved interactome datasets.

2.2 INTRODUCTION

Protein-protein interactions demonstrate functional principles of biological processes because most proteins carry out their cellular functions by interacting with other proteins. The set of all protein interactions within an organism, known as the "interactome", is often represented as a network (Pawson and Nash, 2000; Vidal, 2005). Interactome networks are powerful resources for biologists because they help elucidate the interconnected nature of signaling and communication within cellular systems. It has also been suggested that mechanistic explanations of many human diseases can be obtained by studying alterations to this network (Barabasi et al., 2011; Vidal et al.,

2011). For example, a global guilt-by-association principle has been widely used to predict disease genes by dissecting molecular networks (Oliver, 2000).

However, for an interactome network to be successfully applied in biological studies, it is imperative that it incorporate the intricate structural details of proteins within the network and not simply treat the proteins as mathematical points in a graph-theoretic network (Schuster-Bockler and Bateman, 2008; Wang et al., 2012). Since structure is the basis of protein function (Lahiry et al., 2010), elucidating structural details of interactions can help refine our current understanding of biochemical function from protein-protein interaction networks (Barabasi and Oltvai, 2004).

Here we present INstruct (http://instruct.yulab.org), a comprehensive database of high-quality, three-dimensional (3D), structurally resolved protein interactome networks in human and six widely-studied model organisms. To our knowledge, INstruct is the first online repository containing structurally resolved interaction interfaces between proteins for which no co-crystal structure is available. To accomplish this, we employed an interaction interface inference method (Wang et al., 2012), to structurally resolve interactions based on 37,210 known co-crystal structures in the PDB (Berman et al., 2000). In total, INstruct currently contains 6,585 human, 644 *A. thaliana*, 120 *C. elegans*, 166 *D. melanogaster, 119 M. musculus,* 1,273 *S. cerevisiae*, and 37 *S. pombe* structurally resolved interactions. As a comprehensive database providing structural details not previously annotated in protein interactome networks, INstruct will be an invaluable resource in a wide array of biological research.

2.3 METHODS

2.3.1 Identifying high-quality binary interactions

Binary protein-protein interaction data used to build INstruct was curated from eight major interaction databases – BioGrid (Stark et al., 2011), DIP (Salwinski et al., 2004), HPRD (Keshava Prasad et al., 2009), IntAct (Kerrien et al., 2012), iRefWeb (Turner et al., 2010), MINT (Licata et al., 2012), MIPS (Mewes et al., 2011), and VisAnt (Hu et al., 2009). A binary interaction is a direct, biophysically feasible interaction between two proteins. However, it has been shown that databases that systematically curate protein interactions from the literature could contain erroneous pairs or non-binary interactions (i.e., co-complex associations) (Cusick et al., 2009). Because it is of paramount importance to identify only the high-quality binary interactions for a variety of biological purposes (Das and Yu, 2012), INstruct obtains only binary interactions from the aforementioned databases and filters them rigorously. PSI-MI evidence codes (Hermjakob et al., 2004) for each reported interaction from these databases indicate what experiment was used to ascertain the interacting pair. We retained only those interactions indicated as binary by the supporting evidence codes (Table 2.2) because it is only possible to infer interaction interfaces using assays that can detect direct physical interactions (Wang et al., 2012).

From the aforementioned databases, we compiled a comprehensive list of all publications of high-throughput (HT) experiments (single large-scale studies in which many protein pairs are systematically tested for interactions) that identify protein interactions. Because there are fewer of these studies than conventional small-scale studies, we were able to inspect each publication manually. We only included interactions from those publications whose HT experiments have been verified by traditional orthogonal assays (e.g., co-immunoprecipitation). Additionally, if the authors indicate a subset of their final dataset as "high-quality" or "core," we retained only the interactions

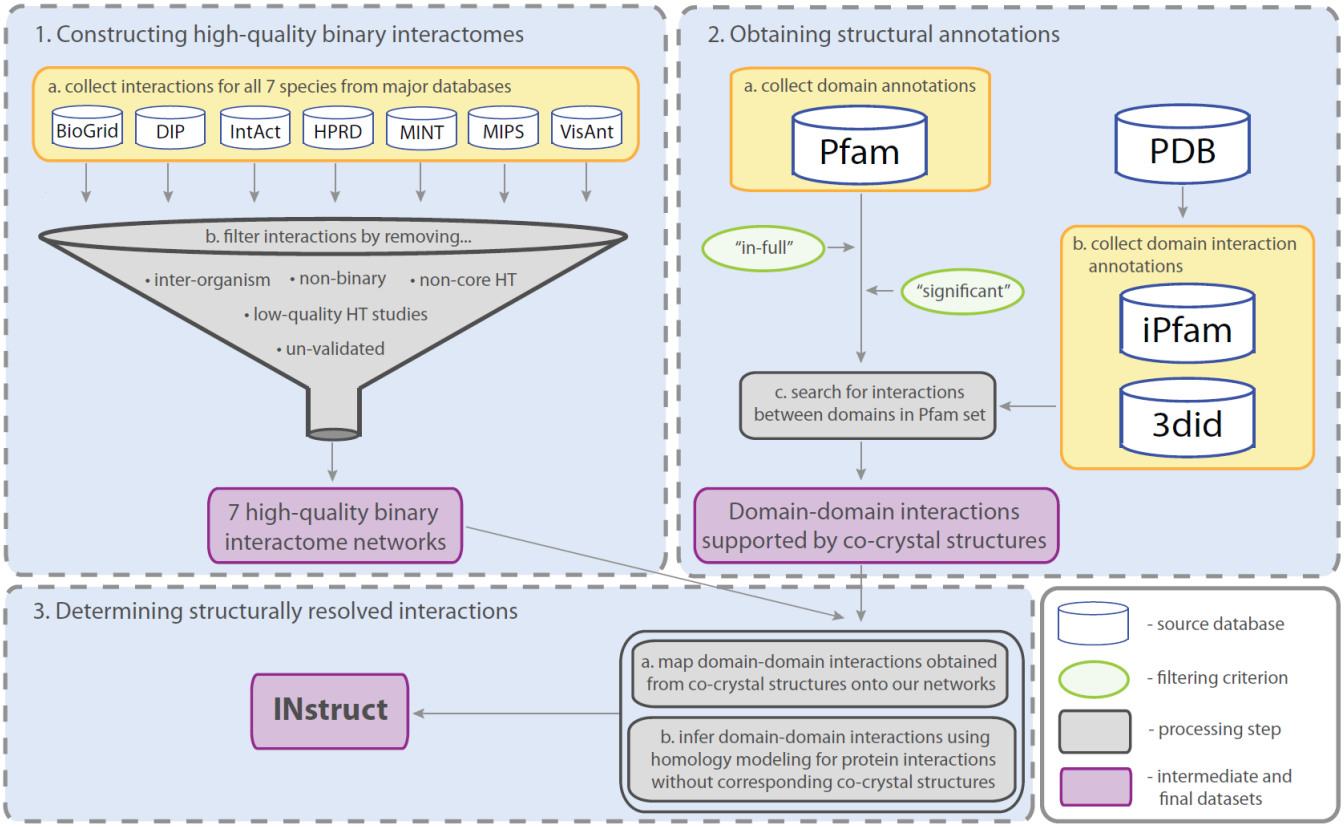

Figure 2.1 A flow-chart showing the sources and three stages of data processing used to create the 3D interactomes in INstruct. (1) Constructing high-quality binary interactomes. Interactomes for each of the seven organisms were created by collecting protein-protein interactions from each of the shown databases. We removed inter-organism interactions, non-binary interactions, interactions from high-throughput (HT) studies that are not a part of the author's high-confidence dataset (core), interactions from low-quality HT studies, and un-validated interactions. (2) Obtaining structural annotations. By collecting high-quality structural annotation data, we produced a set of domain-domain interactions sup-ported by atomic-resolution co-crystal structures. (3) Determining structurally resolved interactions. We employed a method of homology-based interaction interface inference to structurally resolve interaction interfaces for interacting proteins.

meeting the authors' high-confidence criteria. However, for small-scale studies, since it is impossible to manually inspect all papers, we used a well-validated criterion to identify high-quality interactions – it has been shown that interactions supported by two or more publications are of high quality (Das and Yu, 2012; Venkatesan et al., 2009; Yu et al., 2008). Using these criteria, we obtained 61,108 high-quality binary interactions for all seven organisms (full stats in Table 2.1). It should be noted that none of the protein-protein interactions with co-crystal structures are filtered out of INstruct.

2.3.2 Interaction Interface Inference and Validation

In order to add structural resolution to our high-quality binary interactome networks, we leveraged the information in several protein databases. Using protein domain definitions from Pfam (Punta et al., 2012), we identified "Pfam-A" domains, which are both "significant" and "in-full" as defined by Pfam, that also appear in proteins in our high-quality binary interactome networks. To determine the domains mediating the protein-protein interactions in our network, we gathered domain interaction data from 3did (Stein et al., 2009) and iPfam (Finn et al., 2005) which in turn derive their domain-domain interaction evidence from 37,210 existing 3D atomic-resolution co-crystal structures in the PDB. 1,708 protein-protein interactions in our binary interactomes are directly represented by one of these co-crystal structures, in which case it is straightforward to determine where the pair of proteins interact.

For 7,236 protein-protein interactions not supported by direct co-crystal evidence, we applied a tested interaction interface inference method (Wang et al., 2012) to extend the scope of the interaction data provided by 3did and iPfam. (Figure 2.1) For these interactions, we predicted the interface domains based on co-crystal structures of homologous domains for one or both partners. While 3did and iPfam indicate pairs of homologous domains that have been shown to interact in co-crystal structures of pairs of proteins, INstruct is the first source to predict that these domain-domain interactions facilitate protein-protein interactions for which no co-crystal structure exists.

Given a pair of interacting proteins for which there exists a co-crystal structure, we integrated the information from both 3did and iPfam to identify their interacting domains. However, the majority of interactions in INstruct do not fall into this category and therefore must be resolved using our interaction interface inference method.

This approach identifies high-confidence domains catalogued by Pfam in each of two interacting proteins. Pfam curates its Pfam-A set of domain families by constructing seed alignments for each family from a nonredundant, functionally verified set of domain-sequences trusted to be representative of the family. Hidden Markov models (currently based on the package HMMER3) are used to grow each domain family from a set of representative seed domains to include closely homologous domains in other proteins. The resulting family of domains is given a single accession identifier of the form PFXXXXX. To ensure the quality of our method, INstruct only annotates a protein with Pfam-A domains, which have been found "in-full and significant" in the protein, subject to Pfam curation criteria. (Punta et al., 2012)

Once annotated with Pfam-A domains, if proteins in INstruct (which have already been shown to interact in our high-quality binary network) are found to contain domains from families that have been shown to interact in another pair of proteins as indicated by co-crystal evidence in the PDB and catalogued by 3did or iPfam, then the domain pair is also predicted to be an interface which supports the protein interaction for which no co-crystal evidence is available.

Such parsimonious methods of assigning domain interfaces could potentially result in the prediction of domain-domain interactions that do not facilitate a given protein-protein interaction, however we have verified that our predicted domain-domain interactions are high quality. We performed three-fold cross-validation to verify the reliability of the domain-domain interactions inferred by transferring domain interaction interfaces supported by co-crystal structures to interacting protein pairs without co-crystal structures. Into three subsets we split 1,456 human protein interaction pairs that have co-crystal structures. Using two of the subsets at a time as a training set and one as a test set, we predicted domain-domain interactions in the test set using our comparative modeling approach, rather than taking advantage of the co-crystal evidence. We

repeated the procedure using each of the three subsets as the test set in order to ascertain how accurately we could predict interaction interfaces when deprived of co-crystal evidence. We found that we can correctly infer the protein-protein interaction interfaces in over 90% of the 1,456 interaction pairs, indicating high confidence in our method and in the data supplied by 3did and iPfam (Wang et al., 2012).

Though we have demonstrated high confidence in the ability of our method to identify the domains at protein interaction interfaces, it is important to note the inherent difference in resolution available for interfaces determined directly from co-crystal evidence versus those that were inferred using homologous structures. Atomic resolution information is only available for interactions with co-crystal structures; whereas, interaction interfaces inferred from homology are resolved to the level of protein domains. To maintain uniformity, INstruct displays only domain-level information for all interactions. When available, atomic resolution information is easily accessible through direct links to the PDB.

In total, these methods yielded 8,944 3D protein-protein interactions with structurally resolved interaction interfaces. Full network statistics are available in Table 2.1.

2.4 USAGE

A web-based interface is deployed for accessing these interactomes, which includes five basic features: (i) searching for proteins, (ii) retrieval of interaction data, (iii) visualization of protein domains, (iv) creation of custom downloadable datasets, and (v) downloading of entire interactome datasets. A protein in any organism can be queried using its Universal Protein Resource (UniProt) accession ID (UniProt Consortium, 2011) or its corresponding standard gene symbol. Additionally each organism has a single searchable alternate identifier used by a popular organism-specific

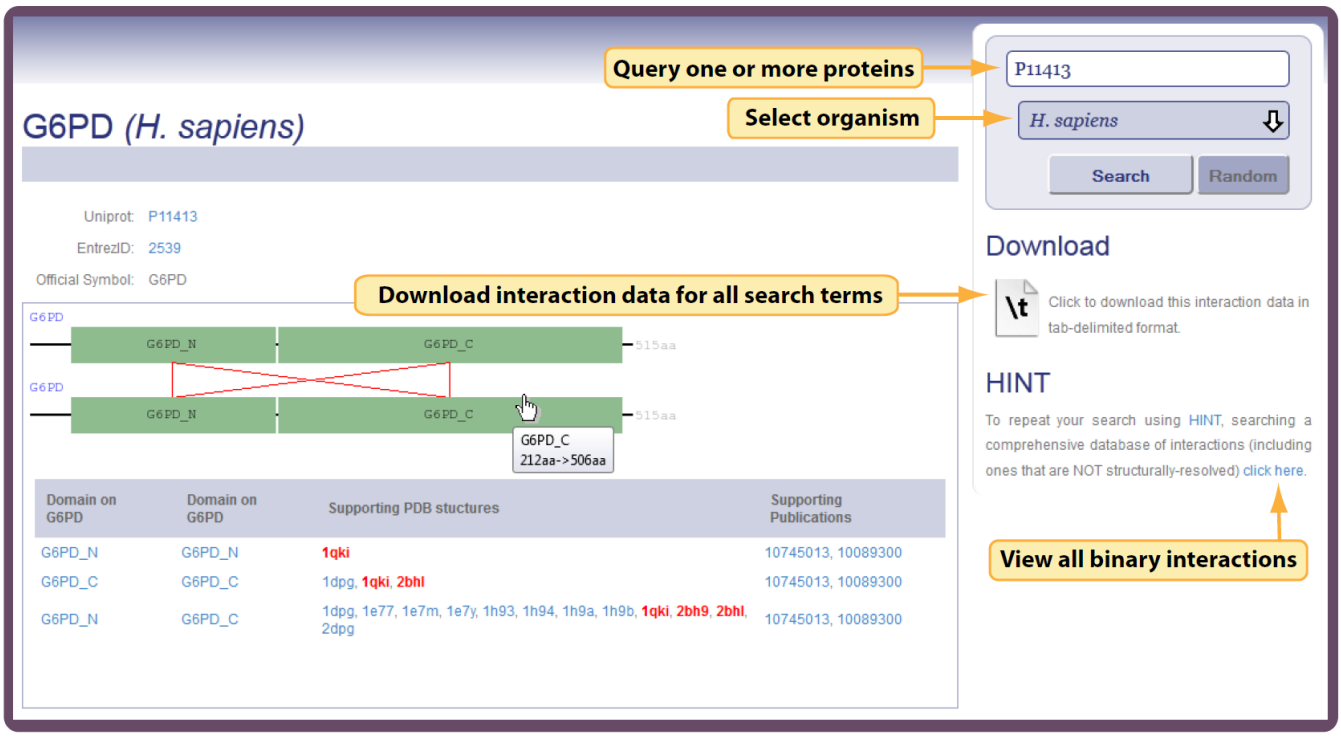

Figure 2.2 Screenshot of the interaction search and retrieval web interface. Shown are the results for a query for the human protein G6PD using its UniProt ID. In this case, the only structurally resolved interaction available is between G6PD and itself. The red edges connecting the domains in this homodimer indicate that the interactions were determined from direct co-crystal evidence; in the table, the PDB structures and publication IDs highlighted in red provide this evidence. PDB structures listed in blue provide homology-based evidence of the domain-domain interaction.

database. The standard query interface (shown in Figure 2.2) accepts multiple simultaneous queries using any combination of the three available identifiers for each organism.

Users are taken directly to the results page if one or more of their queries matches an entry in INstruct. This page shows the results for each query linearly down the page in the order that they appeared in the query. Each matching query returns three types of output: (i) naming information, (ii) schematics showing the domain-level interactions, and (iii) sortable tables providing information about domain-domain interactions. The sidebar always contains the search box, for refinement of search terms, and a dialogue for downloading the complete set of interactions that match all terms in the query.

For each protein that interacts with a query protein, a schematic is shown, displaying the domain architecture of both proteins side-by-side as linear models according the order of their amino acids

(an example is shown in Figure 2.2). Between each pair of proteins, domains that interact are indicated in green and those that do not are indicated in grey. Network edges are drawn between domains that interact on the two proteins, with edges colored in red indicating domain interactions derived directly from co-crystal evidence, and edges in grey indicating domain-domain interactions inferred by homology. Regardless of whether a domain is involved in a structurally resolved interaction, all domain information is interactive and linked to further information provided on the Pfam website.

Interactions involving the queried protein and each of its interacting partners are shown in a table below each schematic and can be sorted by domain on either protein, on the number of publications supporting this interaction, or on the number of supporting PDB structures. PDB structures that provide direct co-crystal evidence for the indicated domain-domain interaction are shown in red, and all other PDB structures provide homology-based evidence and are shown in blue. External references, including Pfam links to each domain in the interacting pair, publication information from PubMed, and PDB structures supporting each interaction, can be accessed by clicking on their corresponding entry in the table.

Additionally, each search produces a tab-delimited text file containing all relevant information about the interactions as shown on the results page, including all available naming conventions and amino acid locations of the interacting domains. A custom dataset can be created simply by searching for up to five proteins at once, separating each term with a delimiter in the search field, or by uploading a file of identifiers on the Downloads page. All valid search terms will be added to the downloadable file and displayed on the results page.

When available, the domain-domain interaction pairs that facilitate the shown protein interactions on the results page will appear in the sidebar on the right side of the results page.

Underneath each domain-domain pair is given a list of proteins that interact via the same domain-domain pair. Proteins in red provide the direct co-crystal evidence for the given domain-domain interaction. In other words, a co-crystal containing each protein in red proves the existence of the domain-domain interaction in the first place and allows us to resolve other protein-protein interactions to include this domain-domain interaction.

2.5 LOOKING AHEAD

INstruct will be a very useful tool in a broad spectrum of fundamental biological research. With the continued growth of data sources, especially available co-crystal structures in the PDB, we expect the coverage of our structurally resolved interactome networks to increase over time (Chandonia and Brenner, 2006). We plan to update INstruct at least once per year to incorporate newly available information and to expand our repertoire of model organisms. As more structural data becomes available, we will build more comprehensive 3D, structurally resolved interactome networks for different organisms.

	H. sapiens	*A. thaliana*	*C. elegans*	*D. melanogaster*	*M. musculus*	*S. cerevisiae*	*S. pombe*
High-quality binary interactions	27,356	12,068	3,928	4,438	1,222	11,936	160
Domain-domain interactions	11,470	825	180	191	169	1,857	52
Structurally-resolved protein-protein interactions	6,585	644	120	166	119	1,273	37
Unique proteins	3,628	454	144	242	130	978	53

Table 2.1 Interactome network statistics.

PSI-MI Accession Code	Assay	PSI-MI Accession Code	Assay
889	Acetylation assay	729	Luminescence based mammalian interactome mapping
14	Adenylate cyclase complementation	231	Mammalian protein protein interaction trap
678	Antibody array	515	Methyltransferase assay
8	Array technology	516	Methyltransferase radiometric assay
872	Atomic force microscopy	671	Monoclonal anitbody
10	Beta galactosidase complementation	77	NMR
11	Beta lactamase complementation	81	Peptide array
809	Bimolecular fluorescence complementation	84	Phage display
12	Bioluminescence resonance energy transfer	434	Phosphatase assay
276	Blue native PAGE	841	Phosphotransfer assay
91	Chromatography	953	Polymerization
16	Circular dichroism	435	Protease assay
17	Classical fluorescence spectroscopy	89	Protein array
27	Co-fractionation	90	Protein complementation assay
807	Comigration in gel electrophoresis	31	Protein cross-linking with a bifunctional reagent
404	Comigration in non denaturing gel electophoresis	424	Protein kinase assay
808	Comigration in sds page	55	Resonance energy transfer
405	Competition binding	227	Reverse phase chromatography
25	Copurification	97	Reverse ras recruitment system
28	Cosedimentation in solution	726	Reverse two hybrid
29	Cosedimentation through density gradient	440	Saturation binding
30	Cross-linking studies	99	Scintillation proyimity assay
406	Deacetylase assay	71	Sizing column
870	Demethylase assay	104	Static light scattering
111	dihydrofolate reductase reconstruction	921	Surface plasmon resonance
38	Dynamic light scattering	107	Surface plasmon resonance
40	Electron microscopy	108	T7 phage display
42	Electron paramagnetic resonance	370	Toy-r dimerization assay
410	Electron tomography	232	Transcriptional complementation assay
411	ELISA	20	Transmission electron microscopy
605	Enzymatic footprinting	397	Two hybrid array
415	Enzymatic study	399	Two hybrid fragment pooling approach
47	Far western	398	Two hybrid pooling approach
48	Filamentous phage display	18	Two-hybrid
49	Filter binding	112	Ubiquitin reconstruction
928	Filter binding	113	Western blot
52	Fluorescence correlation spectroscopy	114	X-ray crystallography
416	Fluorescence microscopy	826	Y ray scattering
53	Fluorescence polarization spectroscopy	115	Yeast display
51	Fluorescence technology	825	Y-ray fiber diffraction
54	Fluorescence-activated cell sorting		
728	Gal4 vp16 complementation		
229	Green fluorescence protein complementation assay		
510	Homogeneous time resolved fluorescence		
858	Immunodepleted coimmunoprecipitation		
19	Immunoprecipitation		
492	In vitro binding		
423	In-gel kinase assay		
859	Intermolecular force		
226	Ion eychange chromatography		
65	Isothermal titration calorimetry		
420	Kinase homogeneous time resolved fluorescence		
66	Lambda phage display		
655	Lambda repressor two hybrid		
369	Ley-a dimerization assay		
67	Light scattering		

Table 2.2 PSI-MI Accession Codes for assays indicating binary protein interactions

2.6 REFERENCES

Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat Rev Genet *12*, 56-68.

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet *5*, 101-113.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Chandonia, J.M., and Brenner, S.E. (2006). The impact of structural genomics: expectations and outcomes. Science *311*, 347-351.

Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M.*, et al.* (2009). Literature-curated protein interaction datasets. Nat Methods *6*, 39-46.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

Finn, R.D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics *21*, 410-412.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C.*, et al.* (2004). The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol *22*, 177-183.

Hu, Z., Hung, J.H., Wang, Y., Chang, Y.C., Huang, C.L., Huyck, M., and DeLisi, C. (2009). VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res *37*, W115-121.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U.*, et al.* (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res *40*, D841-846.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A.*, et al.* (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res *37*, D767-772.

Lahiry, P., Torkamani, A., Schork, N.J., and Hegele, R.A. (2010). Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat Rev Genet *11*, 60-74.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E.*, et al.* (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res *40*, D857-861.

Mewes, H.W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K.F., Stumpflen, V.*, et al.* (2011). MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res *39*, D220-224.

Oliver, S. (2000). Guilt-by-association goes global. Nature *403*, 601-603.

Pawson, T., and Nash, P. (2000). Protein-protein interactions define specificity in signal transduction. Genes Dev *14*, 1027-1047.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res *32*, D449-451.

Schuster-Bockler, B., and Bateman, A. (2008). Protein interactions in human genetic diseases. Genome Biol *9*, R9.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X.*, et al.* (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res *39*, D698-704.

Stein, A., Panjkovich, A., and Aloy, P. (2009). 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. Nucleic Acids Res *37*, D300-304.

Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) *2010*, baq023.

UniProt Consortium (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Research *39*, D214-219.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Vidal, M. (2005). Interactome modeling. FEBS Lett *579*, 1834-1838.

Vidal, M., Cusick, M.E., and Barabasi, A.L. (2011). Interactome networks and human disease. Cell *144*, 986-998.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

CHAPTER 3

BISQUE: locus- and variant-specific conversion of genomic, transcriptomic, and proteomic database identifiers

3.1 ABSTRACT

Biological sequence databases are integral to efforts to characterize and understand biological molecules and share biological data. However, when analyzing these data, scientists are often left holding disparate biological currency—molecular identifiers from different databases. For downstream applications that require converting the identifiers themselves, there are many resources available, but analyzing associated loci and variants can be cumbersome if data is not given in a form amenable to particular analyses. Here we present BISQUE, a web server and customizable command-line tool for converting molecular identifiers and their contained loci and variants between different database conventions. BISQUE uses a graph traversal algorithm to generalize the conversion process for residues in the human genome, genes, transcripts, and proteins, allowing for conversion across classes of molecules and in all directions through an intuitive web interface and a URL-based web service.

3.2 INTRODUCTION

The proliferation of genomic and proteomic databases has helped us organize and understand biological molecules and phenomena, but has left the scientific community using many different naming conventions for the same, or intrinsically related biological entities: genes, proteins, and transcripts (Cunningham et al., 2015; Gray et al., 2015; Pruitt et al., 2014; UniProt-Consortium, 2015). While there are many tools to convert the identifiers themselves, there are insufficient

resources to convert loci and variants annotated in reference to these identifiers (Table 3.1). Due to pervasive sequencing and variant annotation, this has led to a routine burden on biologists to convert from one naming convention to another, and may lead to errors when building upon other labs' research (McCarthy et al., 2014).

Here we present BISQUE (The **Bi**ological **S**e**qu**ence **E**xchange), a multi-interface utility for converting human genomic, transcriptomic, and proteomic loci and variants from their reported form into forms useful for downstream research. BISQUE is an extensible conversion framework deployed as a web server (http://bisque.yulab.org) for user-friendly conversion among the most popular human database identifiers. It is also available as a programmatic web service, downloadable as a customizable standalone application (http://github.com/hyulab/bisque), and importable as a Python module.

3.3 METHODS

BISQUE is designed to track the manifestation of variants in coding regions of human biomolecules. It can trace genomic variants to their effects in higher order molecules (transcripts and proteins) and, conversely, discover the source of proteome aberrations in lower order genomic sequences (genome and genes). As such, BISQUE catalogues positions in molecules that are functional in their transcription or translation across all catalogued molecular classes (genome/genes, transcripts, and proteins). To further functional discovery, BISQUE incorporates two peripheral databases for investigation, dbSNP (Sherry et al., 2001) and the PDB (Berman, 2000), allowing users to quickly determine the relationships between known genomic variants and their potential effects on protein structures.

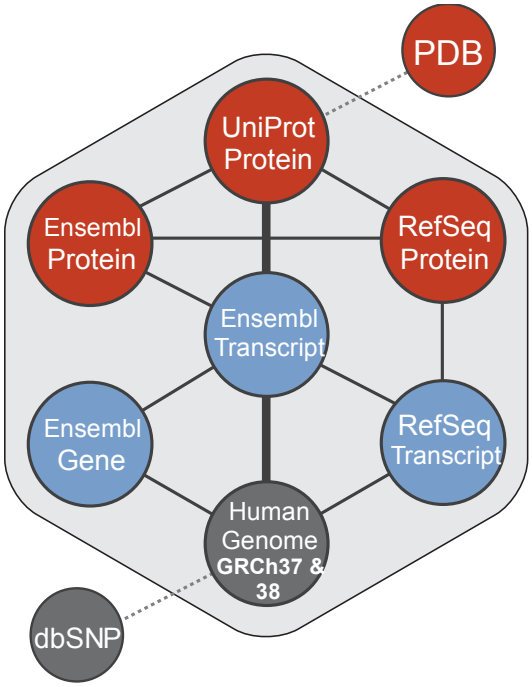

Figure 3.1 The core BISQUE conversion graph, including all possible starting and ending points for a conversion with BISQUE.

The nodes in the BISQUE core conversion graph (Figure 3.1) include the latest human genome builds (GRCh37 and GRCh38) (Benson et al., 2015), Ensembl gene, transcript, and protein (Cunningham et al., 2015), RefSeq transcript and protein (Pruitt et al., 2014), UniProt protein (UniProt-Consortium, 2015), dbSNP (Sherry et al., 2001), and the PDB (Berman, 2000). The connections between these nodes in the graph represent all potential traversals to produce conversions, beginning at an origin node determined by user input and a destination node based on a selection by the user. BISQUE then identifies the optimal path through the conversion graph from the origin to the destination based on the available edges (3.3.5). Conversions are computed stepwise along this path as single conversions between one input node and one output node. At each step, BISQUE first checks if a mapping exists for the input identifier in the selected output database. If a mapping exists, any associated loci and mutations will be converted according to the rules that describe the edge and its direction between the nodes (3.3.6). For instance, when converting from

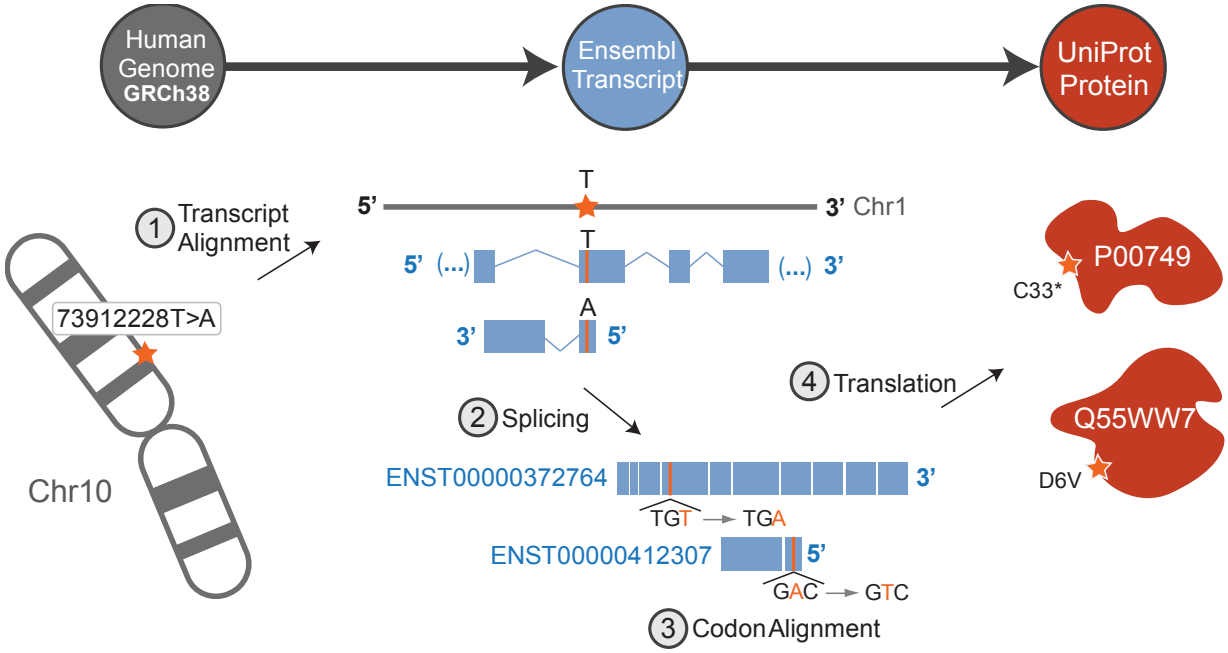

Figure 3.2 The steps for conversion of a genomic variant to UniProt amino acid substitution(s) by traversing a conversion path through Ensembl transcript. The shown genomic variant maps to two Ensembl transcripts, one on the forward strand and one on the reverse. BISQUE uses genomic alignments of transcripts from Ensembl's database, removes introns and non-coding exons (UTRs and alternatively spliced regions), and codon-aligns the result to match amino acids in UniProt proteins.

a genomic locus to a transcript locus, introns must be removed, strand-sense (whether or not the transcript is on the sense (+) or antisense (-) genomic DNA strand) must be taken into account, and variant bases complemented in a transcript annotated in reference to the antisense strand. The output of each stepwise conversion is fed as input to the next stepwise conversion. Figure 3.2 describes an example conversion, showing the input and output of each step required to convert from genomic variant to protein substitution.

3.3.1 Database dependencies

BISQUE is decoupled from runtime dependencies on other databases, and does not owe its allegiance to any one database convention. It must nevertheless rely on mappings provided by these databases to their peers. A useful mapping table contains two parts: the actual correspondence of identifiers in two databases and the biological sequences to which these identifiers correspond.

Together, these two pieces of information provide much of the basis required to convert in both directions between the identifiers.

When both identifier correspondence tables and sequences are available, MySQL tables were constructed to map these values between databases and a template conversion algorithm was customized to convert loci and variants along this new edge in the BISQUE conversion graph (Figure 3.1).

BISQUE stores parsed versions of files at the following URLs in a MySQL database:

UniProt (http://www.uniprot.org/)

- ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping_selected.tab.gz
- ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz
- ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/HUMAN.fasta.gz

Ensembl (http://www.ensembl.org/)

- ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.pep.all.fa
- ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.cds.all.fa
- ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.79.gtf.gz

NCBI (http://www.ncbi.nlm.nih.gov/)

- ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2ensembl.gz
- ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/protein/protein.gbk.gz
- ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/RNA/rna.fa.gz
- ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/GFF/ref_GRCh38.p2_top_level.gff3.gz

- ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b142_GRCh38/ASN1_flat/*
- ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b142_GRCh37p13/ASN1_flat/*

EBI (http://www.ebi.ac.uk/)

- ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/split_xml

We acknowledge that curation practices of BISQUE's source databases may sometimes lead to erroneous results. Since BISQUE is a conversion utility for existing identifiers, it does not attempt to reconcile these errors, which are rare and should be addressed by the databases themselves. However BISQUE will provide the best possible results given the data that is available, and will be updated frequently to incorporate the latest revisions to each external database.

3.3.2 Database updates

Updates of BISQUE can be automated through scripts available in the GitHub repository (http://github.com/hyulab/bisque) that are designed to rebuild all conversion MySQL tables by re-accessing the latest versions of all source databases. However, due to inevitable changes in file formats and FTP server architectures in the source databases, some manual maintenance needs to be performed before each update. Advanced users of BISQUE will be able to perform this maintenance themselves, and we have released all necessary source code to accomplish manual updates. We are also committed to our own routine updates of BISQUE. We intend to update the GitHub update scripts, the downloadable full and lite versions of the software (and associated MySQL tables for download), and the web server twice each year.

3.3.3 Identifiers

BISQUE uses identifiers as reported by their associated databases. Many databases include alternate identifier names, however for internal consistency, BISQUE only uses primary names for each. The following identifiers are currently available in BISQUE's conversion graph (command-line and API handles in parentheses):

Protein identifiers: UniProt (uniprot), The Protein Data Bank (pdb), Ensembl Protein (ensp), RefSeq Protein (refp).

Transcript identifiers: Ensembl Transcript (enst), RefSeq Transcript (reft).

Genomic identifiers: Ensembl Gene (ensg), GRCh37 (grch37), GRCh38 (grch38), dbSNP (dbsnp).

Gene symbols: Gene symbols may be used as input to BISQUE, and will be interpreted as equivalent to the corresponding UniProt protein. This means that associated loci and variants should be made in reference to the UniProt identifier. All other inputs should use systematic database names.

Identifier versions: Some identifiers (most visibly, RefSeq) have associated version numbers appended after the identifiers themselves (i.e., RefSeq transcript NM_000346.3). These version numbers are updated in each new release of the database if the sequence associated with that transcript is updated. BISQUE maintains copies of only the most recent versions of these transcripts, and it is often the case that older transcript versions are forward compatible with the latest database build. Outdated (and therefore potentially erroneous) versions of RefSeq transcripts will be treated as the most recent versions (and notice of the correction will be displayed on the output page).

RefSeq transcripts can also be inputted by the user without any version number indicated, in which case BISQUE will assume the latest available version of that transcript. Most other databases (i.e., Ensembl and UniProt) perform versioning of their identifiers, but the version numbers are hidden from the user. For these, BISQUE uses the latest available versions of all identifiers and their associated sequences and ID mappings to other databases, and does not allow input of version numbers from the user.

3.3.4 Conversion architecture

BISQUE's internal database (built from the sources identified in 3.3.1), primarily consists of two types of tables: (1) mapping identifiers to their residue sequences, and (2) mapping identifiers in one database to those of another, both within and across molecular class (i.e., gene, transcript, protein). Not all identifiers can be mapped to all other identifiers through direct mappings. Rather, BISQUE maintains a requisite number of available conversion tables between identifiers such that all identifiers can be converted to all other identifiers through one or more conversions. This architecture can be depicted as a graph, wherein nodes represent database identifiers and edges represent bidirectional conversions that may take place between identifiers (Figure 3.1). By traversing this conversion graph, BISQUE can produce conversions between any identifiers contained in its database.

3.3.5 Determining optimal paths in the conversion graph

Although BISQUE is capable of traversing any user-defined path in the conversion graph (command-line only), optimal paths from input to output databases are chosen for users of the web server. In defining these paths, we have attempted to avoid spurious results by minimizing the number of

steps in each traversal, avoiding descending and climbing the graph in a single traversal, and minimizing the number of databases used. For instance, when converting from Ensembl Gene to RefSeq Transcript, BISQUE will traverse the path through Ensembl Transcript rather than through the human genome which could yield results in overlapping genes (i.e., on the opposite genomic strand) that are unrelated to the original query. When possible, a single best path is chosen, however in cases where multiple paths are equally optimal based on the aforementioned criteria, BISQUE combines the results from traversing both paths. For example, when converting from Ensembl Transcript to RefSeq Protein, BISQUE traverses from Ensembl Transcript→***Ensembl Protein***→RefSeq Protein and from Ensembl Transcript→***RefSeq Transcript***→RefSeq Protein. Both paths share the same number of steps, number of databases traversed, and neither path traverses both upstream and downstream, making them equally optimal.

3.3.6 Locus- and variant-specific conversion

Conversions in BISQUE are performed stepwise along optimal paths in the conversion graph, with the output from one conversion step being fed as input into the next until an entire conversion path has been traversed from input to output. In cases where multiple outputs are produced from a single conversion step (i.e., due to codon degeneracy), all outputs from that step will be presented to the user (if it is the terminal step in a conversion), or fed as multiple inputs into the next conversion algorithm. Conversion algorithms associated with each edge vary throughout the graph, and must take into account different parameters depending on the type of conversion being made. For instance, converting from a genomic locus to transcript locus follows inherently different rules than converting an amino acid substitution to all potential transcript variants. The primary conversion algorithm types are:

Genomic Locus → Transcript

When a genomic locus is provided, it will be checked to see if it exists within an annotated transcript region. In all transcripts, by default, a locus must be specified relative to the transcript's coding sequence (CDS). However, users may also specify loci in relation to full transcript sequences (including 5’ and 3’ UTRs as annotated by the source database), by selecting the cDNA option through the web interface or command line tool.

To determine the ordinal position of the genomic variant within the transcript, the exons comprising the transcript are joined to create a locus mapping from genomic positions to transcript positions. Genomic positions falling within introns (even those contained between exons in the same transcript) will not map in BISQUE. Both input and output positions are 1-indexed.

Variants given at genomic loci are assumed to be in reference to the sense (+) genomic strand regardless of whether annotated transcripts in that region are derived from the sense or antisense (–) strand. Therefore, when converting variants, BISQUE will return complements of both reference and alternate nucleotides provided as input if the mapped transcript is derived from the antisense strand. If the given reference nucleotide does not match the true reference nucleotide (from the transcript database), the true reference will be assumed and the web server will alert the user that a correction has been made.

Due to performance concerns, it is not possible to convert an entire chromosome without a provided locus to transcripts on the web server (though this can be accomplished through the downloadable command-line tool). Even when a locus is provided, there are many cases in

which several transcripts are derived from the same genomic region (occasionally on opposite genomic strands—see Figure 3.2). BISQUE will report all transcripts when this is the case.

Transcript → Genomic Locus

The reverse procedure can be performed to convert transcripts and associated annotations to the genome. Transcript loci should be provided as 1-indexed positions within the transcript's coding sequence (or full sequence when the cDNA option is selected). Variants should be provided in reference to the raw transcript sequence (single strand mRNA containing ATG start codon), regardless of the polarity of the genomic strand it was derived from. Returned genomic variants will always be reported on the sense (+) strand.

Gene Conversions

Conversions to and from genes follow the same rules as transcript ↔ genomic locus conversions. The only genes currently in the BISQUE core conversion graph are from the Ensembl database, which maintains genomic regions in both GRCh37 and GRCh38 corresponding to their annotated genes. These regions are strict supersets of Ensembl Transcripts, which allows a simple numerical offsets to account for the differences in annotating loci within genes, transcripts, and genomic regions.

A major difference between gene conversions and transcript conversions is that genes are always referenced on the sense (+) genomic strand. Therefore, loci of genes need to be provided as 1-indexed from the first annotated base in the gene, 5'-most on the sense strand, regardless of the polarity of transcripts derived from this gene. Variants annotated in reference to genes should also match the genomic bases of the sense strand.

Please note that the vast majority of intragenic sequence is non-coding, and therefore loci referenced within these regions will not map to other identifiers besides the genome and are not subject to reference-base checking as BISQUE does not store the sequences outside coding regions as a space-saving measure.

Transcript → Protein

Basic conversion of identifiers is handled by mapping tables derived from the database associated with the protein. Loci are converted by separating the coding sequence of the transcript into codons, with codons 1–*n* in the transcript directly aligning to amino acids 1–*n* in the protein. Variants are converted by replacing the reference nucleotide with the given alternate nucleotide and observing the effect on the amino acid encoded by the codon.

If a reference amino nucleotide provided by the user does not match the true reference nucleotide (from the transcript database), the true reference will be assumed and the web server will alert the user that a correction has been made. In rare cases where the reference codon does not match the reference amino acid in the converted protein, the protein reference amino acid will be displayed.

Protein → Transcript

Basic conversion of identifiers is handled by mapping tables derived from the database associated with the protein. Loci are converted by separating the coding sequence of the transcript into codons, and aligning amino acids 1–*n* in the protein directly to codons 1–*n* in the transcript. All possible single-nucleotide variants in the transcript codon (3 possible alternate nucleotides × 3 possible positions = 9 potential variants) are assessed for matches to

the provided alternate amino acid, and all potential nucleotide variants that could produce the provided amino acid substitution are presented to the user. In cases where a single protein can be mapped to multiple transcripts, the entire procedure is completed for each matching transcript, with all results that contain the given locus and could produce the given substitution presented to the user.

If a reference amino acid provided by the user does not match the true reference amino acid (from the protein database), the true reference will be assumed and the web server will alert the user that a correction has been made. In rare cases where the true reference amino acid does not match the codon in the transcript, the protein reference amino acid is assumed true and the given alternate amino acid is matched against all possible single-nucleotide variants within the codon as given by the transcript database.

Transcript ↔ Transcript

Conversions between transcripts are performed by mapping transcript IDs between databases. Loci mapping between transcripts is performed by referring to annotated nucleotide positions within the genome, leveraging the fact that both transcripts have been pre-aligned to the genome. In rare instances where after locus mapping reference alleles at the given positions do not match, the correct reference allele will be shown for both transcripts and variants will not be possible to map.

Protein ↔ Protein

Conversions between proteins are computed using ID mappings from the protein databases (UniProt mappings for conversions involving UniProt proteins and NCBI mappings between

Ensembl and RefSeq). In rare cases wherein protein sequences do not match for matched IDs in different databases, a Needleman-Wunsch-based algorithm, EMBOSS Stretcher (Li et al., 2015), is used to align the two sequences to create a mapping between their loci. These are isolated cases of incompatible identifier versions referring to slightly different protein isoforms, or inter-database inconsistencies over which is considered the 'canonical' isoform of a protein.

3.3.7 Conversion quality filtering options

The following options are available via the BISQUE web interface, API, and command-line tool. These are specifically designed to help users choose the most relevant results in cases of many identifier mappings.

Swiss-Prot only (--swissprot)

When this option is selected, conversion output will be filtered to only include UniProt identifiers with the reviewed, Swiss-Prot designation. Additionally, when converting to identifier types other than UniProt, only those identifiers with direct mappings to SwissProt UniProt identifiers will be included. For instance, when converting a genomic locus to Ensembl transcript position, only transcripts that encode SwissProt UniProt proteins will be returned.

Canonical only (--canonical)

When this option is selected, conversion output will be filtered to only include UniProt identifiers representing canonical isoforms (i.e. with the "-1" designation). Additionally, when converting to identifier types other than UniProt, only those identifiers with direct mappings

to canonical UniProt isoforms will be included. For instance, when converting a genomic locus to Ensembl transcript position, only transcripts that encode canonical UniProt protein isoforms will be returned.

Calculate alignment scores (--quality)

When this option is selected, BISQUE will calculate the alignment scores between sequences for all identifier-based conversions in the conversion path. Alignment scores are calculated as the fraction of identical residues in a sequence alignment for each step in a conversion path for which the two identifiers have directly comparable sequences (i.e. protein-protein, transcript-transcript, gene-transcript, and transcript-protein). These identity scores are then averaged across all steps in a conversion for which alignments could be performed and are reported on a 0-1 scale (1 being identical).

Alignment identity scores are calculated using EMBOSS Stretcher (Li et al., 2015), a Needleman-Wunsch-based algorithm designed to be faster than EMBOSS Needle. All alignments are performed using Stretcher default parameters, protein alignments are performed with the default EBLOSUM62 substitution matrix, and gene and transcript alignments are performed using the default EDNAFULL substitution matrix. For conversion steps that span molecular classes, for instance transcript to protein, nucleotide sequences are converted to amino acid sequences using a standard codon mapping table, and a protein alignment is performed.

Alignment scores for mapping to and from the PDB are performed without Stretcher. Rather, since PDB's only connection to the rest of the conversion graph is through UniProt, the alignment identity score between PDB and UniProt is taken as the number of UniProt residues

contained in the PDB structure based on SIFTS residue mappings (Velankar et al., 2013) divided by the total number of residues in the UniProt protein. As there are sometimes many PDB structures with varying degrees of coverage of the same UniProt protein, these alignment scores are very useful to identify which PDB structures contain the largest number of the original UniProt residues.

It should be noted that sequence alignment is never used to determine identifier mappings; these are always determined using mapping tables provided by the source databases (although alignments are used to determine residue mappings for protein-protein alignments as discussed in 3.3.6). The alignment score simply provides a metric to assess the compatibility of these identifier mappings.

3.4 RESULTS

3.4.1 Usage

BISQUE is available in several forms to increase its usability for a variety of applications. For small queries (up to 1000 conversions at a time), the web server is recommended. For larger queries and systematic integration, the web service or command-line application may be more appropriate.

The BISQUE web server is both a frontend to the conversion engine and a portal to learning more about available web functions and accessing BISQUE in its other forms (discussed below). Queries can be submitted either one at a time or in batch (by pressing the '+' button) through the interactive form on the home page. Batch queries may be manually entered into a form or uploaded as a text file in a variety of formats, including the Variant Call Format (VCF). Conversion results are presented in a table which is exportable in several popular formats including CSV, JSON, and XML.

BISQUE also provides simple programmatic access to its conversion engine through URL-based queries that produce tab-delimited, plain text output. By passing query parameters through CGI fields, users may rapidly access BISQUE's conversion capabilities in their own scripts (examples using both Python and Perl are provided on the About page), or incorporate BISQUE conversions into publicly available web servers and databases.

The BISQUE conversion engine is encapsulated in a separate, downloadable command-line application. Two versions of this tool are available for download through the BISQUE web server—a lite and a full version. The full version download includes all MySQL tables required for mapping so that a user may install BISQUE without any external dependencies. The lite version contains all of the same functionality as the full version, but relies on an internet connection to access data from the public BISQUE MySQL database.

The command-line application is open source, written in Python, and importable as a Python module, providing users with another avenue to incorporate BISQUE into their scripts. This also gives more advanced users the ability to modify BISQUE to meet their own needs, for instance to expand the coverage of database identifiers available for conversion by adding and removing edges and nodes from the conversion graph. For most new databases this is achievable through included utilities and documentation, and may not require modifying the code.

3.4.2 Database & conversion scope

The set of databases over which BISQUE can perform conversions was selected to best represent the most popular conventions of current large datasets (Abecasis et al., 2012; Forbes et al., 2011; Fu et al., 2013). Many datasets are released with annotations to several database conventions, making it highly likely that BISQUE will be useable for the vast majority of applications. We have chosen to

focus on biomolecules within these databases that can be converted completely within our conversion framework (i.e., proteins, mRNA transcripts, and coding regions of genes and genomes). Such conversions are more suited to an all-by-all conversion framework as they can begin and end at any node within the conversion network. Furthermore, this ensures a small database footprint, improving BISQUE's speed and portability.

However, we realize that research takes place outside of coding regions of our chosen core set of databases. To ensure that BISQUE is more generally useful, we have designed it to be extensible beyond this core set of identifiers. Experienced programmers will appreciate the ability to *fork* BISQUE on GitHub (http://github.com/hyulab/bisque) and the scripts and documentation we have made available to aid in the integration of more databases (nodes) within the BISQUE conversion graph.

Since BISQUE focuses on mutations with functional effects in all molecular classes, we have also included two databases that fall outside of our core conversion graph, but have high functional relevance: The Protein Databank (Berman, 2000) (PDB) and The Single Nucleotide Polymorphism Database (Sherry et al., 2001) (dbSNP). Maintaining our curation of molecules and regions with functional effects throughout the entire conversion graph, we curate missense SNPs in dbSNP (~ 1.3 million) and regions of PDB structures that map to UniProt proteins (UniProt-Consortium, 2015; Velankar et al., 2013).

3.4.3 Comparison with other conversion utilities

Part of the reason, we believe, that there are few off-the-shelf options for locus and variant-specific conversion is that many online data analysis tools compute these conversions on a case-by-case basis. For instance, predictors of variant function, which are not designed to act as standalone

conversion tools, still must maintain some *under-the-hood* conversion capability if they are to be user-friendly and accept data annotated in variety of conventions. One popular functional SNP predictor, PolyPhen-2 (Adzhubei et al., 2010), allows a variety of inputs, but ultimately converts all of these to UniProt in order to perform the multiple sequence alignments that contribute to its classifier. Other functional annotation tools such as snpEff (Cingolani et al., 2012) and VAT (Habegger et al., 2012) also convert downstream starting from chromosome variants in VCF files to variants in transcripts and proteins (Table 3.1).

However the conversion frameworks underlying these tools and others are limited to converting in one direction, from many types of inputs into a single type of output, not freely among all identifiers. For instance, popular variant effect predictor VEP (McLaren et al., 2010), while it provides built-in conversion capability to provide a better user experience, is limited to a nucleotide-centric form of variant conversion. Since it assesses variant effects at the nucleotide level, this is not an issue for the functionality of VEP, however, it means that it cannot handle conversion from one protein identifier to another (i.e. Ensembl Protein to UniProt) when there is more than one possible nucleotide mutation encoding an amino acid substitution. For a non-nucleotide-centric conversion utility, such as BISQUE, these conversions are handled without passing through a nucleotide identifier at all.

BISQUE also provides the ability to perform reverse mappings of variants (i.e. amino acid substitutions to nucleotide mutations). While this functionality may be less often used than forward mapping, it is still addresses some practical scientific needs. For instance, reverse mappings are required for designing PCR primers to introduce specific amino acid substitutions in a protein expression system by altering the nucleotides of the encoding transcript. Such a task is not suitable for nucleotide-centric variant annotation software given the potential multiplicity of nucleotide

mutations giving rise to the same amino acid substitution. BISQUE simply returns all potential nucleotide substitutions, which a researcher may filter based on their criteria for primer design.

Many biological resources do require external conversion for custom queries. For instance, a biologist may be interested in determining the precise 3D locations of UniProt amino acids affected by non-synonymous SNPs in X-ray crystallographic protein structures in the PDB (Berman, 2000) (to assess for proximity to other variants, binding domains, etc.). Alternatively, a biologist having identified a residue of interest in a PDB structure may wish to determine its coding location in the human genome. The PDB does not currently have an inbuilt method to convert between genomic variants and amino acid substitutions, nor should it as its developers cannot possibly anticipate all of the potential uses of their data. BISQUE removes the onus of developing conversion utilities from databases like the PDB, which are likely to have little interest in maintaining conversion utilities in addition to their own data.

3.5 DISCUSSION

While there are some tools available to perform limited conversion of positions and variants associated with database identifiers (Table 3.1), these often exist as coupled conversion-analysis tools, which lack the all-to-all conversion capability of BISQUE. Furthermore, variant conversion is typically coupled with much more computationally expensive functions, such as variant annotation. What is lacking in these tools is modularity—a core design principle of complex systems that allows functional components to be repurposed for other tasks. BISQUE embraces modularity in its own internal structure through a generalized conversion framework, which enables the expansion of its conversion capability. BISQUE is also a module itself, with the ability to perform multiple conversions for a scripting task or to act as the conversion engine for a database or analysis tool. In

this way, BISQUE will be very valuable to the scientific community as it publicizes vital biological infrastructure.

Resource	URL	Type	Input Vehicle / Type	Outputs	Variant- or locus-specific conversion?
UniProt (UniProt-Consortium, 2015)	http://www.uniprot.org	Proteomic Database	Web form & service; UniProtKB, Ensembl, RefSeq, etc.	UniProtKB*	No
DAVID (Huang da et al., 2009)	http://david.abcc.ncifcrf.gov	Bioinformatics Suite	Web form & service; UniProtKB, Ensembl, RefSeq, Microarray (Agilent, Affy), etc.	Same as inputs	No
Synergizer (Berriz and Roth, 2008)	http://llama.mshri.on.ca/synergizer/translate	Database ID Conversion	Web form; UniProtKB, Ensembl, RefSeq, PDB, etc.	Same as inputs†	No
PICR (Cote et al., 2007)	http://www.ebi.ac.uk/Tools/picr	Protein ID Conversion	Web form; UniProtKB, Ensembl, RefSeq, etc.; amino acid sequences	Protein Identifiers: UniProtKB, PDB, etc.	No
Ensembl Biomart (Cunningham et al., 2015)	http://www.ensembl.org/biomart	Bulk Ensembl Data Retrieval	Web form & API; Ensembl Identifiers	UniProt, RefSeq, Genebank, etc.	No
PolyPhen-2 (Adzhubei et al., 2010)	http://genetics.bwh.harvard.edu/pph2	Functional Annotation	Web form; Proteins and SNPs; amino acid sequences	UniProtKB, Annotation	Yes‡ to UniProtKB
SnpEff (Cingolani et al., 2012)	http://snpeff.sourceforge.net	Functional Annotation	Downloadable Tool; VCF files§	Annotated VCF files§	Yes‡ to Ensembl transcript/gene and amino acid subs
VAT (Habegger et al., 2012)	http://vat.gersteinlab.org/index.php	Functional Annotation	Downloadable Tool; VCF files§	Annotated VCF files§	Yes‡ to Ensembl transcript/gene and amino acid subs
Condel (Gonzalez-Perez and Lopez-Bigas, 2011)	http://bg.upf.edu/fannsdb/query/condel	Functional Annotation	Web form; UniProtKB, Ensembl, GRCh37	Text tables among all input databases	Yes, among noted databases
VEP (McLaren et al., 2010)	http://www.ensembl.org/Tools/VEP	Functional Annotation	Web form and API; UniProtKB, Ensembl, RefSeq, etc.	Same as inputs	Yes, except for protein-protein conversions
Variant Annotation Integrator (Karolchik et al., 2014)	https://genome.ucsc.edu/cgi-bin/hgVai	Functional Annotation	Web form; Genomic variants as pgSnp or VCF	Text table of UCSC gene identifiers	Yes, only from genome to UCSC gene identifiers

Table 3.1. Survey of available tools for database identifier conversion and variant or locus-specific conversion coupled with functional annotation.

*UniProt can output as many identifier types as can be input, but either the output or the input identifier must be UniProtKB.
†Varies depending on chosen input type.
‡Conversion only available with added overhead of functional annotation and only in one direction to noted outputs.
§Format for storing genetic variants (chromosomal loci and variations).

3.6 REFERENCES

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2015). GenBank. Nucleic Acids Res *43*, D30-35.

Berman, H.M. (2000). The Protein Data Bank. Nucleic Acids Research *28*.

Berriz, G.F., and Roth, F.P. (2008). The Synergizer service for translating gene, protein and other biological identifiers. Bioinformatics *24*, 2272-2273.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92.

Cote, R.G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., and Hermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. BMC Bioinformatics *8*, 401.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.*, et al.* (2015). Ensembl 2015. Nucleic Acids Res *43*, D662-669.

Forbes, S., Bindal, N., Bamford, S., Cole, C., Kok, C., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research *39*, 50.

Fu, W., O'Connor, T., Jun, G., Kang, H., Abecasis, G., Leal, S., Gabriel, S., Rieder, M., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet *88*, 440-449.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A. (2015). Genenames.org: the HGNC resources in 2015. Nucleic Acids Res *43*, D1079-1085.

Habegger, L., Balasubramanian, S., Chen, D.Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., and Gerstein, M. (2012). VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics *28*, 2267-2269.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res *42*, D764-770.

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y.M., Buso, N., and Lopez, R. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res *43*, W580-584.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.B., and Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. Genome Med *6*, 26.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069-2070.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M.*, et al.* (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res *42*, D756-763.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research *29*, 308-311.

UniProt-Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

Velankar, S., Dana, J., Jacobsen, J., van Ginkel, G., Gane, P., Luo, J., Oldfield, T., O'Donovan, C., Martin, M.-J., and Kleywegt, G. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research *41*, 9.

CHAPTER 4

A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human

4.1 ABSTRACT

Here, we present FissionNet, a proteome-wide binary protein interactome for *S. pombe*, comprising 2,278 high-quality interactions, of which ~50% were previously not reported in any species. FissionNet unravels previously unreported interactions implicated in processes such as gene silencing and pre-mRNA splicing. We developed a rigorous network comparison framework that accounts for assay sensitivity and specificity, revealing extensive species-specific network rewiring between fission yeast, budding yeast, and human. Surprisingly, although genes are better conserved between the yeasts, *S. pombe* interactions are significantly better conserved in human than in *S. cerevisiae*. Finally, cross-species interactome mapping demonstrates that coevolution of interacting proteins is remarkably prevalent, a result with important implications for studying human disease in model organisms. Overall, FissionNet is a valuable resource for understanding protein functions and their evolution.

4.2 INTRODUCTION

Proteins function primarily by physically interacting with other proteins. Gain or loss of these interactions within an organism can modulate protein functions and disease states (Sahni et al., 2015; Wei et al., 2014). The importance of protein interactions to our understanding of fundamental biological processes has spurred the mapping of protein interactome networks for several

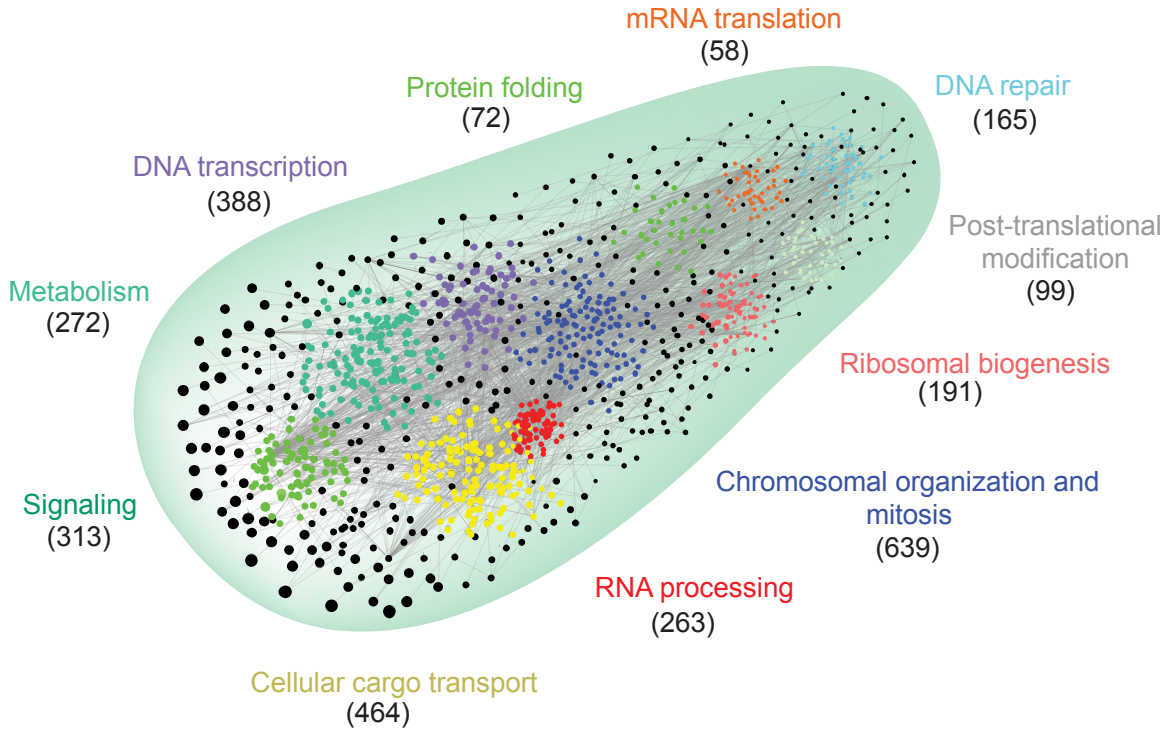

Figure 4.1 Network representation of FissionNet. Proteins are color-grouped based on PomBase GO slim categories. The number of FissionNet interactions per group is indicated. (Figure by T. Vo)

organisms (Arabidopsis Interactome Mapping Consortium, 2011; Giot et al., 2003; Rolland et al., 2014; Stelzl et al., 2005; Yu et al., 2008). However, the budding yeast *Saccharomyces cerevisiae* remains the only eukaryotic organism for which a high-coverage binary protein interactome has been mapped by systematic interrogation of pairwise combinations of all proteins in triplicate (Yu et al., 2008). Here, we present FissionNet, a high-coverage proteome-wide protein interactome network generated for the fission yeast *Schizosaccharomyces pombe*.

We compared FissionNet with the only other proteome-scale eukaryotic interactomes available (>50% of all protein pairs screened), the interactome networks of *S. cerevisiae* and human. Surprisingly, we find that FissionNet is more similar to the human network than it is to that of *S. cerevisiae*. Furthermore, among interactions involving conserved proteins, there is significant species-specific rewiring that is not completely determined by overall sequence similarity of

orthologs. Instead, we identify several other determinants of interaction conservation, including local network constraints and conservation of interacting protein domains. Also, by comparing FissionNet with the proteome-wide interactome of *S. cerevisiae*, we are able to ascertain how gene duplication events influence the functional evolution of paralogs.

4.3 RESULTS

4.3.1 A proteome-wide high-coverage binary protein interactome map of *S. pombe*

To generate a proteome-wide interactome network for *S. pombe*, which we call FissionNet, we systematically tested all pairwise combinations of proteins encoded by 4,989 *S. pombe* genes (corresponding to >99% of all *S. pombe* coding genes) using our high-quality yeast two-hybrid (Y2H) assay, the same pipeline that we used to generate the budding yeast and human interactome networks (Yu et al., 2008; Yu et al., 2011). Extensive screenings in triplicate (a total of ~75 million protein pairs) yielded 2,278 interactions between 1,305 proteins, of which 2,130 (93.5%) have not been previously reported in *S. pombe* (Figure 4.1) (Das and Yu, 2012). Furthermore, FissionNet contains 1,034 interactions that have not been reported between orthologs in any other species before. Of these, 142 interactions involve *S. pombe* proteins that both have human orthologs, but at least one does not have a *S. cerevisiae* ortholog and, hence, cannot be studied in *S. cerevisiae*. Thus, FissionNet provides a valuable repertoire of biological insights.

To assess the sensitivity and specificity of our Y2H assay (Yu et al., 2008), we constructed a positive reference set (PRS) consisting of 93 well-validated *S. pombe* interactions from the literature and a negative reference set (NRS) of 168 random *S. pombe* protein pairs that are not known to

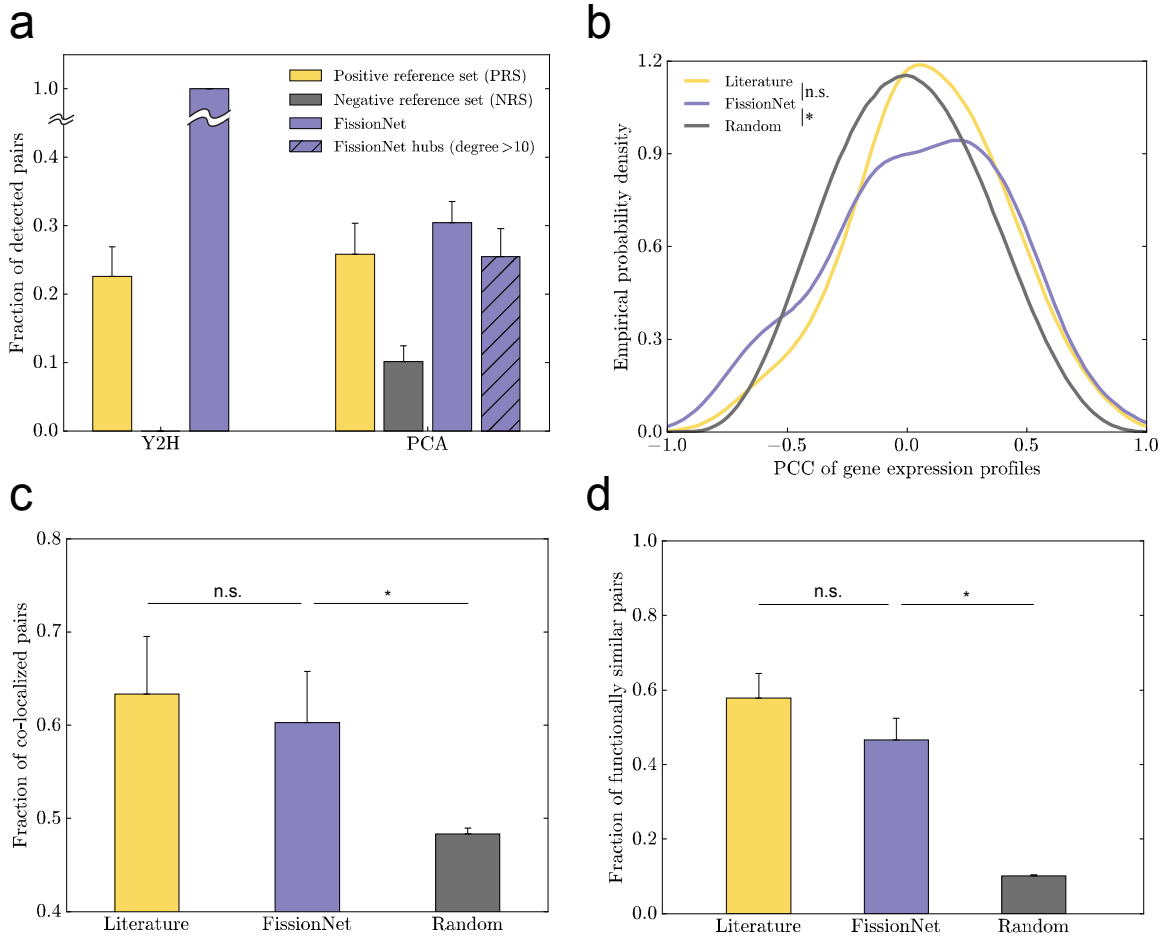

Figure 4.2 (a) Y2H and PCA detection rates of the PRS, NRS, FissionNet, and FissionNet hub interactions. (b) Pearson correlation coefficient (PCC) distribution of gene expression profiles of interacting and all random protein pairs. (c) Enrichment of co-localized protein pairs. (d) Enrichment of protein pairs sharing similar functions. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

interact in the literature and whose orthologs in other species are also not known to interact. We performed Y2H and protein complementation assay (PCA) (Das et al., 2013; Yu et al., 2008) to test what fraction of the PRS, NRS, and a random sample of 220 FissionNet interactions can be detected using orthogonal methods (Figure 4.2a). We found that the detection rates of the PRS and FissionNet interactions are indistinguishable from each other and are significantly higher than that of the NRS (Figure 4.2a; >15% difference in detection rates between the PRS and NRS for both assays, $P<10^{-3}$, *Z* test). The robust validation rates of FissionNet interactions by an orthogonal assay confirm the high quality of the network. Furthermore, although it has been speculated that Y2H interactions

involving proteins with many interaction partners (hubs) could be of low quality (Bader et al., 2004), we found that the validation rate by PCA of hub interactions is the same as the overall PCA validation rate for FissionNet (Figure 4.2a; P=0.34, Z test), confirming that FissionNet interactions involving hubs are of high quality.

Biological relationships between interacting proteins in FissionNet were assessed by measuring similarities in protein localization, functional annotations, and expression profiles. We found that FissionNet interactions are significantly enriched for protein pairs that are co-localized, functionally similar, and encoded by coexpressed genes relative to random expectation (Figures 4.2b to 4.2d; P<0.05 in all three cases using a KS test for coexpression and Z test for co-localization and functional similarity). Furthermore, the enrichment of these interactions for all three categories is similar to that of literature-curated binary interactions. These results confirm that FissionNet interactions are functionally relevant *in vivo*.

4.3.2 Comparative network analyses reveal species-specific conservation of interactions

High-quality protein interactome networks have previously been reported in budding yeast (Yu et al., 2008) and human (Rolland et al., 2014). A fundamental question is how protein-protein interactions have evolved and whether this trend mirrors gene-level evolution. From sequence-based phylogenetic analyses, the two yeasts are less divergent from each other than either yeast is from human (Figure 4.3a) (Sipiczki, 2000). Additionally, the two yeasts share a greater fraction of protein-coding genes than either yeast does with human (Figure A.2).

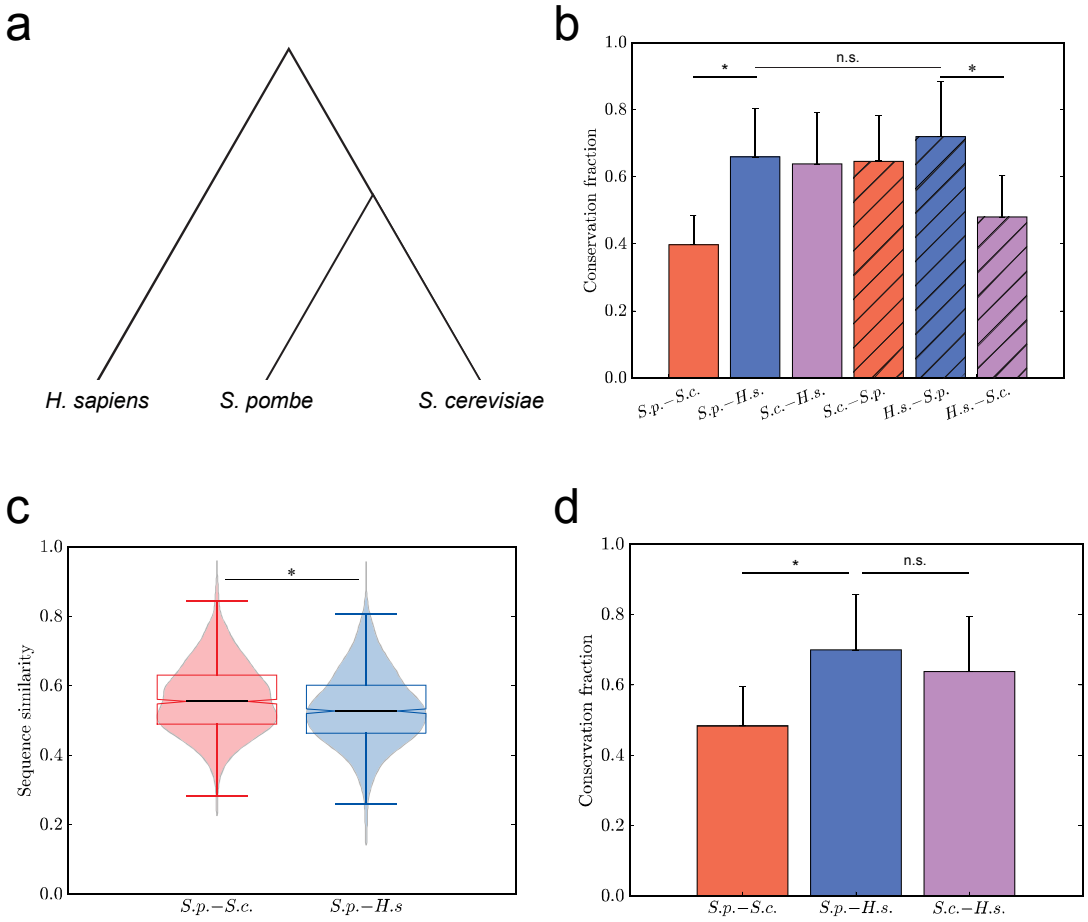

Figure 4.3 (a) Sequence-based phylogeny dendrogram of S. pombe (S.p.), S. cerevisiae (S.c.), and human (H.s.). (b) Interaction conservation between reference-query species. (c) Sequence conservation for ortholog pairs that could be conserved between S.p.-S.c. and S.p.-H.s. (d) Interaction conservation between reference-query species for proteins that are conserved in all three species. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant. Abbreviations are S. pombe (S.p.), S. cerevisiae (S.c.), and human (H.s.).

To calculate interaction conservation, we considered only those interactions that have the potential to be conserved, *i.e.*, the two interacting proteins in the reference species have orthologs in the other species. However, directly calculating the overlap between sets of interactions obtained from the literature would be erroneous because currently available interactomes are incomplete and are derived from assays with varied and often unreported false positive and false negative rates (Yu et al., 2008). Therefore, to accurately estimate the underlying interaction conservation fractions, we required interactomes of all species to be derived from the same experimental assay. Since interactomes in budding yeast (Yu et al., 2008) and human (Rolland et al., 2014) have been

generated using our version of Y2H, we were able to compare FissionNet to these interactome networks to measure the observed extent of interaction conservation. We developed a rigorous Bayesian framework that incorporates both the false positive and false negative rates of our Y2H assay to estimate the underlying interaction conservation fraction from the observed fraction for each pair of species (see Supp. Section A.5). Surprisingly, we find that interaction conservation follows a completely different trend from gene conservation (Figures 4.3b). While only ~40% of *S. pombe* interactions are conserved in *S. cerevisiae* (of the 1,331 interactions where both proteins have *S. cerevisiae* orthologs and were pairwise retested using our Y2H assay), ~65% of *S. pombe* interactions are conserved in human (of the 652 interactions where both proteins have human orthologs and were pairwise retested using our Y2H assay) (Figure 4.3b; $P=1.4\times10^{-4}$, *Z* test). However, when using budding yeast as the reference species, the fraction of conserved interactions is as high in fission yeast as in human, comparable to the fraction conserved between fission yeast and human (Figure 4.3b). We were able to recapitulate these results using interaction datasets generated by other assays (Figure A.3; >1.5 fold difference between fission yeast interactions conserved in budding yeast and human; $P<10^{-3}$ in all cases, *Z* test). Thus, our results suggest that a large fraction of interactions are conserved between human and *S. pombe*, but have been lost specifically in the *S. cerevisi*ae lineage.

One possible explanation for these surprising results is that fission yeast proteins that are conserved in human could have higher overall sequence similarity than those that are conserved in budding yeast. However, we find that proteins in interactions that have the potential to be

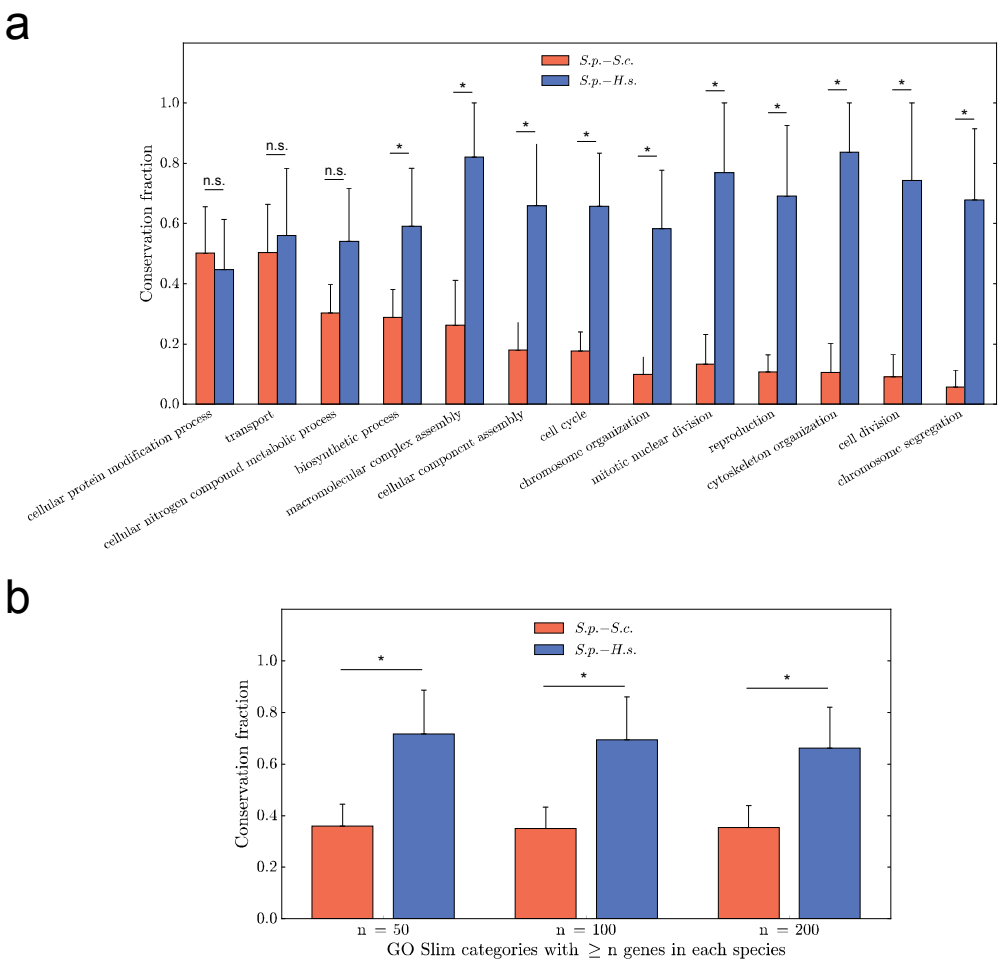

Figure 4.4 (a) Interaction conservation in GO Slim categories with at least 50 interactions. (b) Interaction conservation among GO Slim categories that are conserved in all three species. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant. Abbreviations are S. pombe (S.p.), S. cerevisiae (S.c.), and human (H.s.).

conserved based on orthology are actually slightly more similar in sequence between the two yeasts than between *S. pombe* and human (Figure 4.3c; $P<10^{-5}$, *U* test).

Another possibility is that the observed difference primarily arises from interactions involving proteins that are conserved between fission yeast and human but lost in budding yeast. To test this, we first focused on proteins that are conserved in all 3 species. We still find that ~20% more interactions are conserved between *S. pombe* and human as compared to between the two yeasts (Figure 4.3d; $P<0.05$, *Z* test).

We next explored the conservation of interactions involved in various biological processes as defined by the Gene Ontology (GO) (Ashburner et al., 2000). We find wide variation in species-specific interaction conservation among different processes (Figures 4.4a and A.4). We show that *S. pombe* interactions are more conserved in human than *S. cerevisiae* for 10 out of 13 GO Slim categories containing ≥50 interactions (Figure 4.4a; $P<0.05$, as marked, *Z* test). The same trend is observed with GO Slim categories containing ≥30 or ≥75 interactions (Figures A.4a and A.4c). Some of these categories, such as "chromosomal organization", "chromosome segregation", and "cell cycle", are far better conserved in human than in *S. cerevisiae*, and accordingly *S. pombe* has been used as a model organism for studying these processes (Wood et al., 2002). Furthermore, considering GO Slim categories that are well conserved in all three species (using cutoffs of ≥50, 100, and 200 genes annotated per species), we find that the conservation of *S. pombe* interactions in these core biological processes is also higher in human than in *S. cerevisiae* (Figure 4.4b; $P<10^{-3}$, *Z* test). Overall, these results suggest that insights gained from FissionNet may be widely applicable to the study of human biology across many important cellular processes.

4.3.3 Determinants of interaction conservation

Previous studies have shown that increased protein sequence similarity facilitates conservation of protein interactions (Matthews et al., 2001). Indeed, we also found a positive correlation between sequence similarity of proteins and the fraction of their associated interactions conserved between *S. pombe* and human or *S. cerevisiae*, demonstrating a proteome-scale dependence of protein sequence and function (Figure 4.5a; $R^2_{S.p\text{-}H.s}=0.948$ and $R^2_{S.p\text{-}S.c}=0.976$). However, protein interaction conservation is not completely dependent on overall sequence similarity, as we find many instances of conserved interactions involving proteins with low overall sequence similarity (<40%) with their

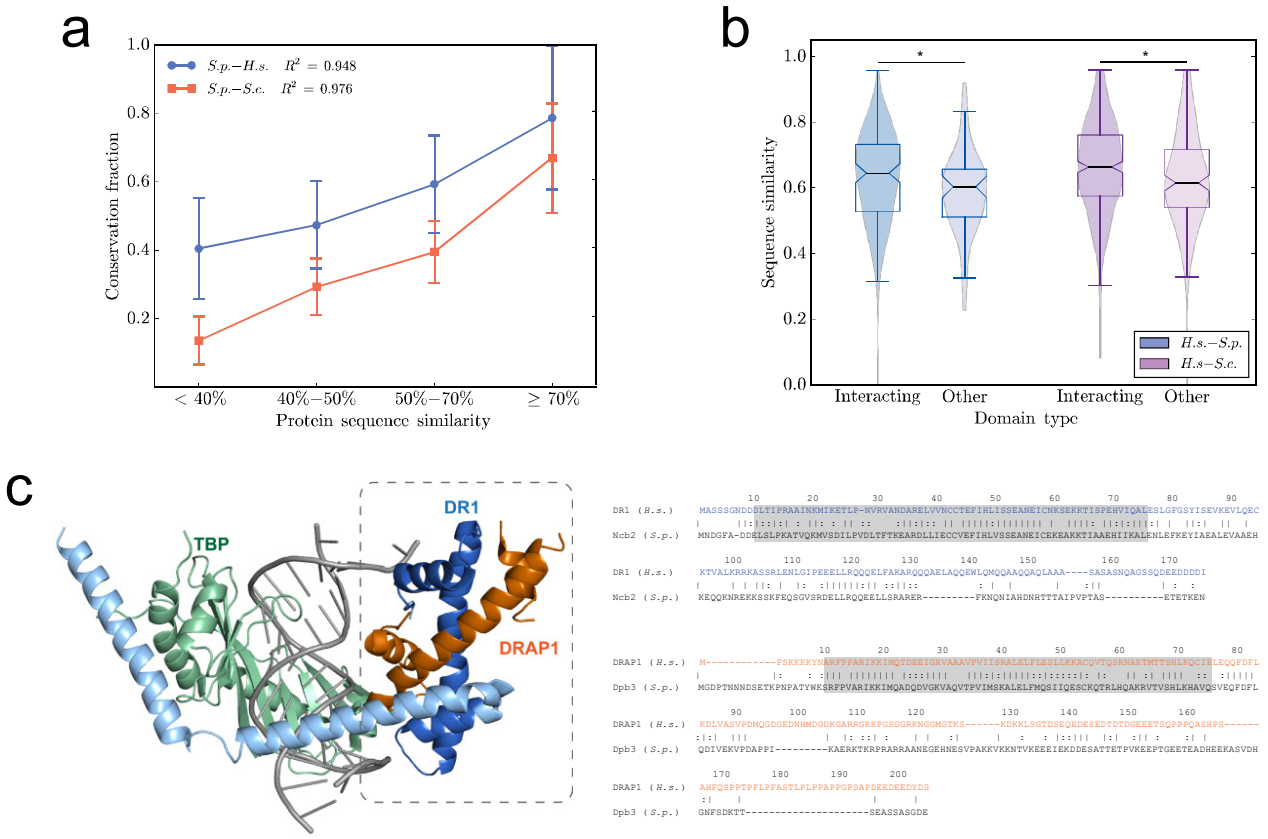

Figure 4.5 (a) Interaction conservation as a function of overall protein sequence similarity. (b) Sequence similarity within protein interaction domains and other domains for interactions conserved between yeasts and human. (c) Crystal structure of human DR1-DRAP1. Boxed region highlights interaction domains. Gray shaded regions denote aligned interaction domain sequences. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant. Abbreviations are S. pombe (S.p.), S. cerevisiae (S.c.), and human (H.s.).

orthologs (Figure 4.5a; 40% and 13% of 116 interactions conserved in human and 196 interactions conserved in *S. cerevisiae*, respectively). To investigate whether certain highly conserved domains in these proteins play an important role in interaction conservation, we inferred protein interaction domains from co-crystal structures of 124 human interactions conserved in *S. pombe* and 293 conserved in *S. cerevisiae*. We find that the sequence similarity within protein interaction domains tends to be higher than in other domains for interactions conserved between fission yeast and human (Figure 4.5b; 7.0% higher, *P*=0.012, *U* test). For instance, the human DR1-DRAP1 heterodimer is orthologous to the protein pair Ncb2 and Dpb3 in *S. pombe.* While the overall sequence similarity of the orthologs is quite low (0.58 and 0.51, respectively), the interaction is

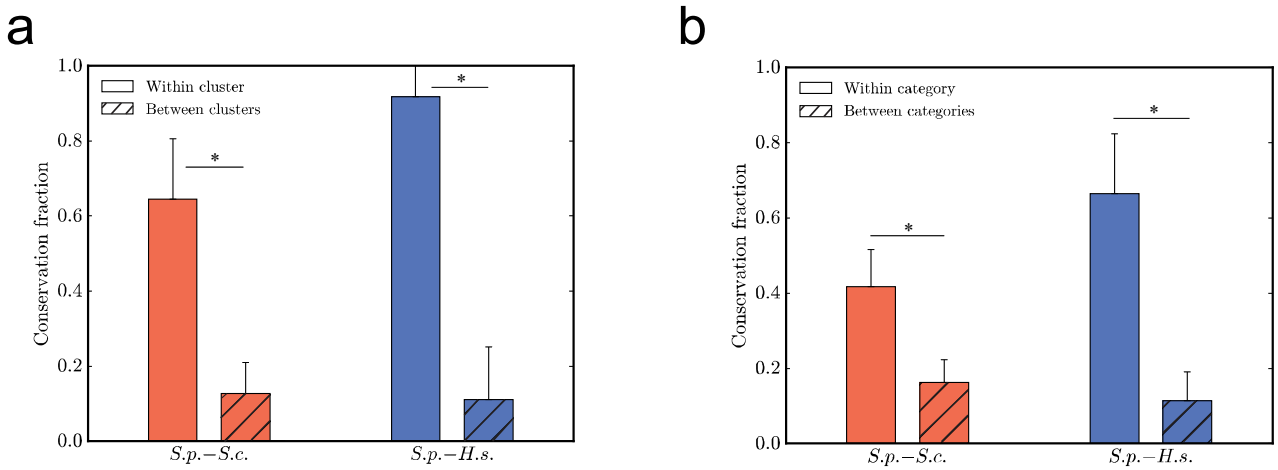

Figure 4.6 (a) Interaction conservation within and across topological clusters. (b) Interaction conservation within and across GO categories. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant. Abbreviations are S. pombe (S.p.), S. cerevisiae (S.c.), and human (H.s.).

conserved in fission yeast. Moreover, we also find that the proteins can interact with the orthologs of their native interaction partner. Based on a crystal structure of the human DR1-DRAP1 complex, we were able to determine the interaction domains of these proteins (Figure 4.5c) (Kamada et al., 2001). The sequence similarity within these domains in DR1 and DRAP1 with their fission yeast orthologs is 0.78 and 0.80, respectively, while the conservation outside of these interaction domains is only 0.45 and 0.38. Thus, the basis for this high degree of functional conservation is likely dependent on the interaction domains.

Strikingly, interaction conservation is nearly three times higher between *S. pombe* and human than between the two yeasts at low levels of overall sequence similarity (Figure 4.5a; at <40% similarity, P=0.030, Z test). As sequence similarity approaches 100%, interaction conservation converges. Therefore, for the vast majority of interactions corresponding to proteins with lower sequence similarity to their orthologs, our results strongly suggest that species-specific factors, independent of overall protein sequence similarity, influence conservation of protein-protein interactions.

We then sought to explore other factors that could explain the basis of interaction conservation. First, we used ClusterOne (Nepusz et al., 2012) to detect topological protein clusters in FissionNet. We find that intra-cluster FissionNet interactions are >3 times more likely to be conserved in both budding yeast and human than inter-cluster interactions (Figure 4.6a; $P<0.05$ for both organisms, Z test). Next, we examined biological processes defined by GO (Ashburner et al., 2000) and observed the same trend (Figure 4.6b; $P<10^{-3}$ for both organisms, Z test). Using genetic interactions, it has been earlier hypothesized that while individual functional modules are conserved, inter-modular connectivity could be rewired across evolution (Roguev et al., 2008). In this study, we provide direct molecular level evidence on a proteome scale that while interactions within modules tend to be conserved across evolution, the cross-talk among these modules changes significantly from one species to another.

4.3.4 Coevolution of conserved interactions revealed by cross-species interactome mapping

To further dissect the nature of conserved interactions, we implemented a cross-species interactome mapping approach to determine the prevalence of coevolution. We consider an interaction to be coevolved when its proteins have evolved in a coordinated manner to maintain the interaction in different species, but have developed incompatible binding interfaces with orthologs of their partners. To determine whether conserved interactions are intact or coevolved, we test by Y2H whether a protein in one species can interact with the ortholog of its interacting partner in another species. If the cross-species interaction can occur, the interaction is intact, otherwise it is coevolved between the two species (Figure 4.7a). For example, through our cross-species mapping, we discovered that interactions of farnesyltransferase subunit Cwp1 with other subunits Cpp1 and

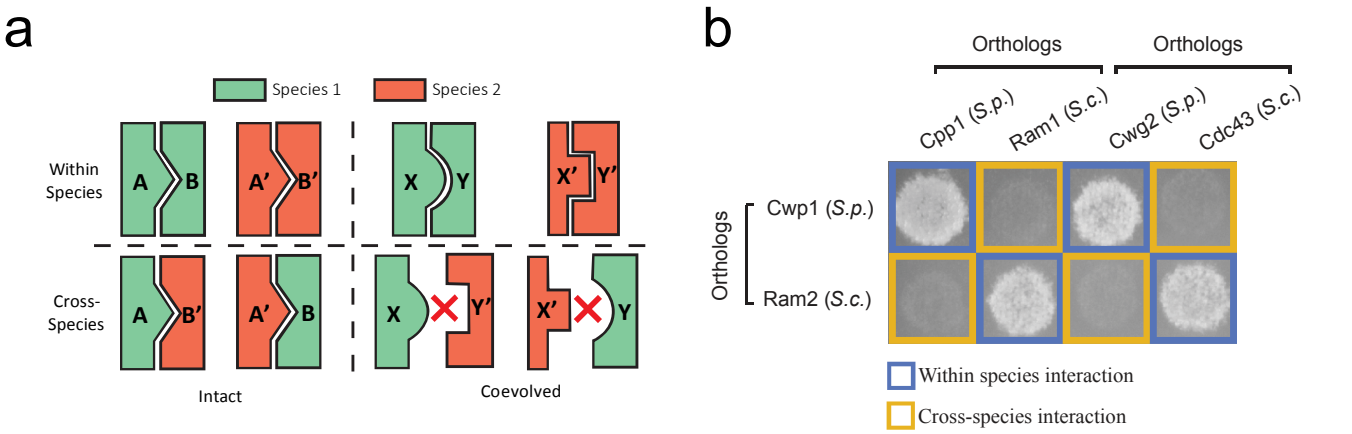

Figure 4.7 (a) Schematic representation of conserved protein interactions that are either intact or coevolved. (b) Within- and cross-species Y2H detects coevolved interactions.

Cwg2 have coevolved between *S. pombe* and *S. cerevisiae*; Cwp1 cannot interact with either Ram1 or Cdc43, *S. cerevisiae* orthologs of Cpp1 and Cwg2, respectively (Figure 4.7b). A previous study showed that expression of Cwp1 cannot complement a non-functional mutant of its *S. cerevisiae* ortholog, Ram2 (Arellano et al., 1998). This suggests that Cwp1, although conserved between *S. pombe* and *S. cerevisiae* at the gene level, has evolved incompatible interaction interfaces with other farnesyltransferase subunits in *S. cerevisiae* and is thus unable to reconstitute an active enzyme complex.

It is known that evolution in protein folds is essentially the result of many random mutation events (Lockless and Ranganathan, 1999). However, since only a small fraction of changes that occur via random drift will satisfy the pairwise constraints necessary for interaction conservation, coevolution at the residue level only occurs at a few specific sites and is relatively rare (Talavera et al., 2015). Surprisingly we find that coevolution at the interaction level is not uncommon: ~33% and 50% of conserved interactions between *S. pombe* and *S. cerevisiae* or human are coevolved, respectively (Figure 4.8a). This shows that even among conserved interactions, only a few key alterations at important binding sites can make the cross-species interactions incompatible and the interactions coevolved. Thus, these sites are critical to protein binding and subsequent function, and

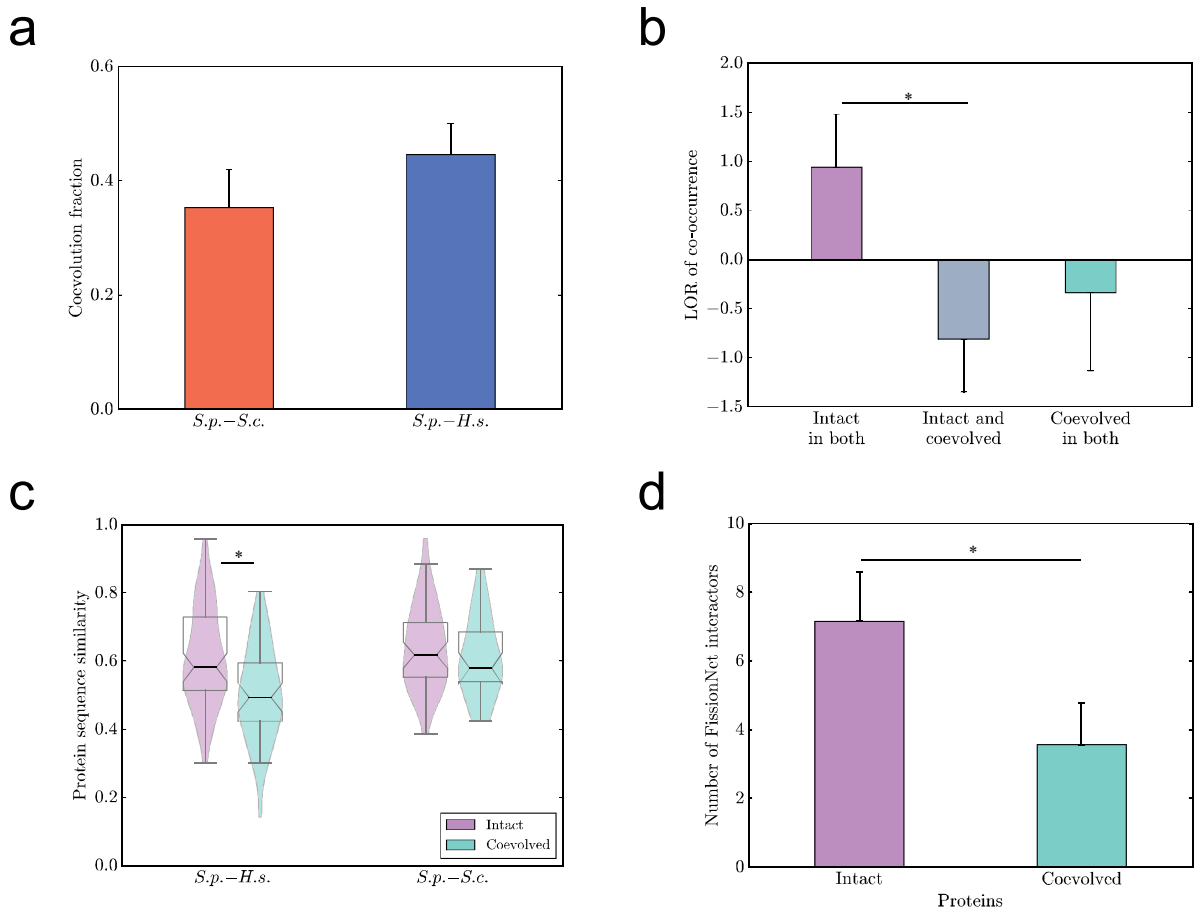

Figure 4.8 (a) Fraction of S.p. interactions that are coevolved with respect to S.c. or human (H.s.). (b) Log odds ratio of co-occurrence of intact and coevolved interactions between S.p.-S.c. and S.p.-H.s. (c) Overall protein sequence similarity of S.p. proteins involved in intact or coevolved interactions. (d) Number of interactors for proteins involved in intact or coevolved interactions. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

changes at these sites alter protein interactions in a manner analogous to a single amino acid change disrupting protein interactions in human disease (Wang et al., 2012; Wei et al., 2014).

Among interactions for which we were able to determine coevolution status, we found that the likelihood for an interaction to be intact between *S. pombe*-*S. cerevisiae* and *S. pombe*-human is significantly higher than random expectation, while the likelihood for an interaction to be intact for one species pair and coevolved for the other species pair is significantly lower (Figure 4.8b; difference in log odds ratio=1.7, *P*=0.022, *Z* test). Thus, these intact interactions are likely involved in functions that have remained unchanged among yeasts and human throughout evolution.

We then investigated potential factors that could determine whether an interaction is intact or coevolved with respect to another species. We find that overall sequences of proteins involved in intact interactions tend to be better conserved across species than sequences of proteins in coevolved interactions (Figure 4.8c; 18.0% higher, P=2.1×10^{-4} for *S. pombe*-human, *U* test). High sequence conservation may indicate higher levels of evolutionary constraint existing within the local network neighborhood of a given interaction. In fact, we find that proteins involved only in intact interactions have twice the number of interactors as compared to proteins involved in only coevolved interactions (Figure 4.8d; P=1.1×10^{-3}, *U* test), suggesting that the added evolutionary constraint of maintaining many interacting partners may prevent the coevolution of two interacting proteins. Finally, we find that the most highly evolutionarily correlated inter-protein residue pairs in coevolved interactions are significantly more correlated than top residue pairs in intact interactions (Figure A.5; P<10^{-10}, *U* test; see Supp. Section A.15), suggesting that the maintenance of coevolved interactions involves compensatory changes at the amino acid residue level.

4.3.5 Implications of FissionNet for the study of human disease

We explored the relevance of FissionNet to human disease by considering the context of known human disease mutations from HGMD (Stenson et al., 2014) within proteins of the human interactome conserved in *S. pombe*. We find that among human interactions conserved in either *S. pombe*, *S. cerevisiae*, or both, ~40% of inter-protein pairs of disease mutations cause the same disease (Figure 4.9a). This is significantly higher than in human interactions that are not reported to be conserved in either yeast or cannot be conserved in either due to lack of protein orthologs (Figure 4.9a; P<10^{-10} for all pairwise comparisons, *Z* test). Based on these results, mutations that

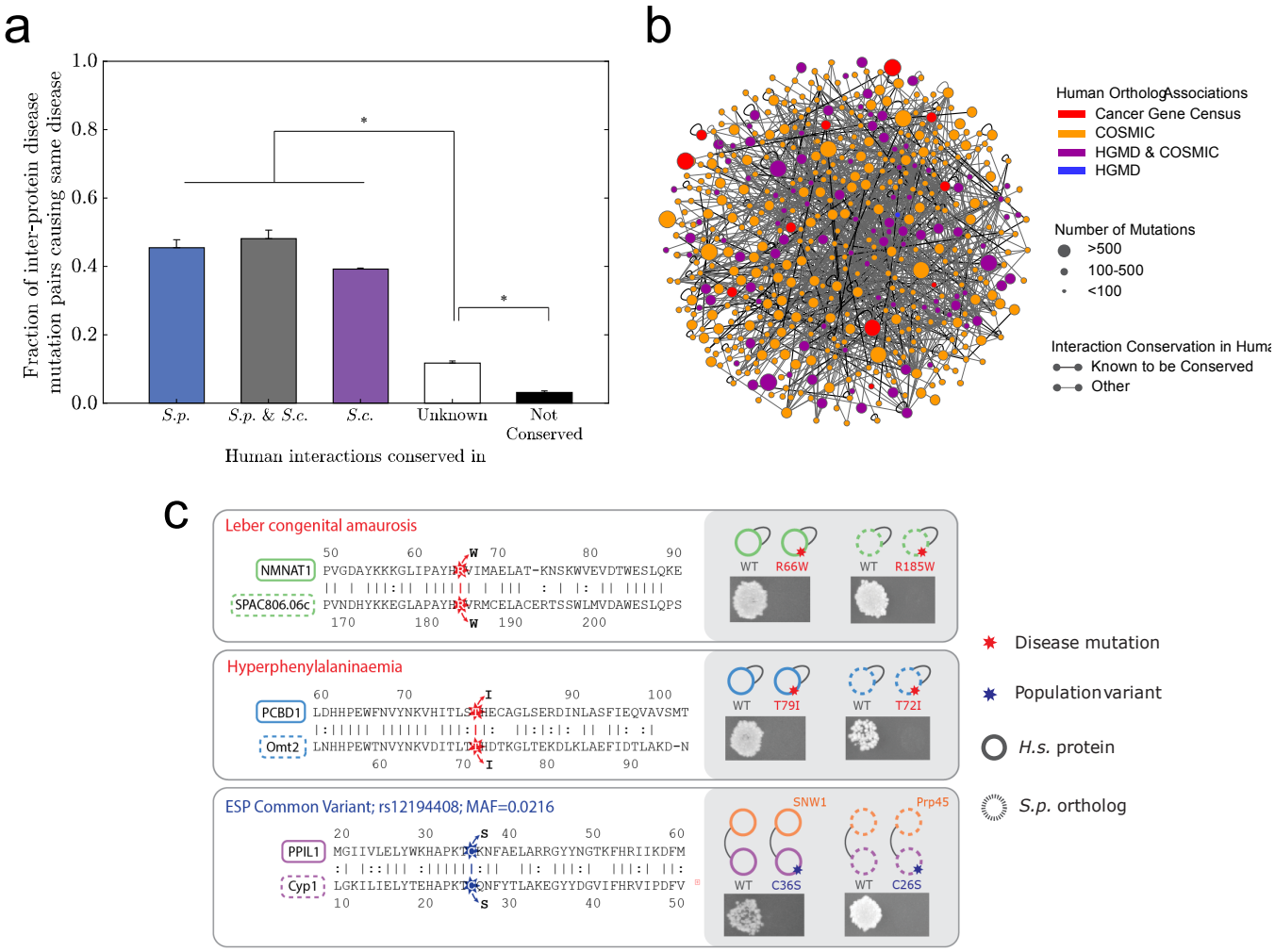

Figure 4.9 (a) Fraction of inter-protein HGMD mutation pairs that cause the same disease in human interactions with regard to their conservation status in S. pombe and S. cerevisiae. (b) Largest connected subcomponent of FissionNet wherein all proteins have human orthologs with known germline disease or somatic cancer-associated mutations. (c) Impact of human disease mutations and a population variant on intact interactions between human and fission yeast. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

break specific protein-protein interactions to cause diseases may be overrepresented among interactions conserved in model organisms. From a global network view, FissionNet may be highly relevant to the study of human disease based on the large portion of *S. pombe* interactions in which both proteins have human orthologs with known germline disease or somatic cancer-associated mutations (Figure 4.9b; 902 interactions) (Forbes et al., 2015; Stenson et al., 2014).

To demonstrate the plausibility of studying specific human disease mutations using FissionNet, we explored whether human disease mutations that disrupt human interactions intact in *S. pombe*

also disrupt the corresponding interactions of the fission yeast orthologs. We focused on three examples: two Mendelian disease variants (Stenson et al., 2014) that disrupt the human NMNAT1-NMNAT1 and PCBD1-PCBD1 interactions and one population variant from the Exome Sequencing project (Fu et al., 2013) that disrupts the human SNW1-PPIL1 interaction. We find that introducing these human protein residue changes into their *S. pombe* orthologs also disrupts the fission yeast interactions (Figure 4.9c). These results indicate that cross-species interactome mapping enables investigation of whether interaction interfaces are altered at the molecular level between model organisms and human, a finding with potentially far-reaching implications for the study of protein function and human disease.

4.4 DISCUSSION

By comparing FissionNet to protein networks in budding yeast and human, we have shown that the molecular bases for interaction conservation among orthologous proteins are complex and different from those that underlie gene conservation. This is highly relevant to the use of the two yeasts as model organisms as there are functions that can be better studied using fission yeast. We find that divergence across species is not completely dictated by sequence level changes, suggesting that rewiring of interactomes plays an important role in species evolution. Additionally, our finding that proteins in a significant fraction of conserved interactions have undergone coevolution to maintain interactions has major implications for studies reliant on the expression of human proteins in model organisms to identify functional mechanisms (Tardiff et al., 2013).

Our analyses focus on budding yeast, fission yeast, and human, as they are the only three eukaryotic organisms for which we have proteome-scale interactome networks using our version of the Y2H assay (>50% of all protein pairs screened). Once more interactome networks are

systematically generated in other species, using assays with measured sensitivity and specificity, the comparative network analysis framework established in this study can be readily applied to further elucidate the extent and nature of the evolution of protein functions across many species.

4.5 REFERENCES

Arabidopsis Interactome Mapping Consortium (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Arellano, M., Coll, P.M., Yang, W., Duran, A., Tamanoi, F., and Perez, P. (1998). Characterization of the geranylgeranyl transferase type I from Schizosaccharomyces pombe. Mol Microbiol *29*, 1357-1367.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Bader, J.S., Chaudhuri, A., Rothberg, J.M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol *22*, 78-85.

Das, J., Vo, T.V., Wei, X., Mellor, J.C., Tong, V., Degatano, A.G., Wang, X., Wang, L., Cordero, N.A., Kruer-Zerhusen, N.*, et al.* (2013). Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. Sci Signal *6*, ra38.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S.*, et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res *43*, D805-811.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E.*, et al.* (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727-1736.

Kamada, K., Shu, F., Chen, H., Malik, S., Stelzer, G., Roeder, R.G., Meisterernst, M., and Burley, S.K. (2001). Crystal structure of negative cofactor 2 recognizing the TBP-DNA transcription complex. Cell *106*, 71-81.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295-299.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res *11*, 2120-2126.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods *9*, 471-472.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.O., Hayles, J.*, et al.* (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. Science *322*, 405-410.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212-1226.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y.*, et al.* (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647-660.

Sipiczki, M. (2000). Where does fission yeast sit on the tree of life? Genome Biol *1*, REVIEWS1011.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet *133*, 1-9.

Talavera, D., Lovell, S.C., and Whelan, S. (2015). Covariation Is a Poor Measure of Molecular Coevolution. Mol Biol Evol *32*, 2456-2468.

Tardiff, D.F., Jui, N.T., Khurana, V., Tambe, M.A., Thompson, M.L., Chung, C.Y., Kamadurai, H.B., Kim, H.T., Lancaster, A.K., Caldwell, K.A.*, et al.* (2013). Yeast reveal a "druggable" Rsp5/Nedd4 network that ameliorates alpha-synuclein toxicity in neurons. Science *342*, 979-983.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N., *et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., *et al.* (2002). The genome sequence of Schizosaccharomyces pombe. Nature *415*, 871-880.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., *et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X., *et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat Methods *8*, 478-480.

CHAPTER 5

mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome

5.1 ABSTRACT

A new algorithm and web server, mutation3D (http://mutation3d.org), proposes driver genes in cancer by identifying clusters of amino acid substitutions within tertiary protein structures. We demonstrate the feasibility of using a 3D clustering approach to implicate proteins in cancer based on explorations of single proteins using the mutation3D web interface. On a large scale, we show that clustering with mutation3D is able to separate functional from non-functional mutations by analyzing a combination of 8,869 known inherited disease mutations and 2,004 SNPs overlaid together upon the same sets of crystal structures and homology models. Further, we present a systematic analysis of whole-genome and whole-exome cancer datasets to demonstrate that mutation3D identifies many known cancer genes as well as previously underexplored target genes. The mutation3D web interface allows users to analyze their own mutation data in a variety of popular formats and provides seamless access to explore mutation clusters derived from over 975,000 somatic mutations reported by 6,811 cancer sequencing studies. The mutation3D web interface is freely available with all major browsers supported.

5.2 INTRODUCTION

A hallmark of the genomic era has been the application of whole-genome and whole-exome sequencing to the study of genetic disease, especially cancer. This effort has led to the development of new statistical methods (Hodis et al., 2012; Lawrence et al., 2013; Sjöblom et al., 2006), which

have identified many potential genomic targets of interest by combing the deluge of data produced by large cohort studies. While these methods have been largely successful in identifying genes with previously unknown roles in tumorigenesis, we have yet to fully realize the promised boon to therapeutic development—although the list of potential disease-causing and driver mutations has grown, the list of approved therapeutics has remained largely static (Das et al., 2014a).

Although the underlying causes of this lag are complex, they can at least be partially attributed to the level of resolution of current methods, which typically identify potentially functional genes based on mutation frequencies at the level of whole genes (Cancer Genome Atlas, 2012; Lawrence et al., 2014; Vucic et al., 2012; Wood et al., 2007). However, many genes carry out a diverse set of functions (pleiotropy), the derangement of any one of which may be sufficient to cause cancer. Further, disruption of different functions of the same gene often lead to clinically distinct types of cancer (Hanahan and Weinberg, 2011; Muller and Vousden, 2013). Finally, even when a specific gene has been identified as being potentially involved in tumorigenesis, researchers may have little idea as to which of its functions has been disrupted. All of these challenges facing current methodologies make it difficult to develop targeted therapeutic strategies.

Here we present mutation3D, an algorithm and web server (http://mutation3d.org) designed to identify somatic cancer-causing genes by leveraging the structure-function relationships inherent in their protein products. In tumorigenesis, mutations are selected that confer a competitive advantage to pre-cancerous cells. Since many mechanisms of tumorigenesis involve alterations to protein function, and protein function is determined by protein structure, tumorigenically selected driver mutations may localize to positions that will affect protein structures. Therefore, mutations causing the same cancer type in a cohort of patients may form clusters (or hotspots) in regions of protein structures wherein alterations confer a competitive

advantage to tumor cells by disrupting specific protein functions. For instance, mutations localized at interaction interfaces may disrupt protein complexes or transient interactions (Wang et al., 2012; Wei et al., 2014), and mutations localized in the hydrophobic core may destabilize the protein entirely (Das et al., 2014b; Kucukkal et al., 2015; Nishi et al., 2013; Petukh et al., 2015).

Recent studies have begun to leverage structure-function relationships in proteins to predict cancer gene targets by searching for nonrandom distributions of mutations in protein crystal structures (Kamburov et al., 2015) and enrichment across protein domains (Miller et al., 2015). We present the first tool to identify and visualize individual clusters within protein structures. Furthermore, we also provide an option to search for clusters in homology models, expanding our coverage of the human proteome more than three-fold (Supp. Section A.1). Through an intuitive, freely available web interface, researchers can use mutation3D to inspect clusters of amino acid substitutions in an interactive molecular viewer to determine whether to follow up with the target based on its structural features. Furthermore, mutation3D can analyze data from whole-genome sequencing (WGS; throughout used to also include whole-exome) studies to perform cluster analysis of variants at the level of the structural proteome.

5.3 METHODS

5.3.1 mutation3D clustering algorithm

The algorithm underlying the mutation3D web interface is complete-linkage (CL) clustering (Sørensen, 1948), a hierarchical clustering method in which clusters first comprise single elements and are then merged with nearest neighboring clusters or unassigned elements until a single cluster comprises all elements. Notably, the clusters found by complete-linkage clustering, as opposed to single-linkage clustering (Sneath, 1957), are assured to have a diameter less than or equal to a

specified linkage distance, which results in tight well-defined clusters. Because of this property, this method can also be referred to as furthest-neighbor clustering, since the dissimilarity of elements within a cluster is determined by the distance between the two elements furthest from each other in n-dimensional space.

In our implementation of this classic machine learning algorithm, we cluster the three-dimensional locations of the α-carbons of those amino acids whose codons contain missense mutations. The coordinates of all atoms within proteins were derived from both PDB structures and structural models (Pieper et al., 2011) based on PDB entries covering proteins either in part or in full. For any given protein, many overlapping models may be available from either or both sources. mutation3D will invariably use entries from the PDB when they are available, as these experimentally determined crystal structures are considered to be a 'gold standard' in structural biology. To increase structural coverage of the proteome, the user may also select a subset of homology-based models to include, based upon several quality metrics available via the Advanced Query page (Supp. Section A.2). Once a set of PDB structures and structural models has been established for a single protein, mutation3D attempts to cluster amino acid substitutions on all models separately, and reports any model or experimentally determined structure in which a cluster has been found. In our analyses we consider it sufficient to implicate a protein in cancer if any of its models are found to contain a cluster.

Some whole proteins or regions of proteins may not have been crystallized or modeled to-date. Owing to the lack of structural coordinates in these regions, we would be unable to identify clusters of mutations. There are some cases in which a single genomic mutation may give rise to defects in distinct proteins, in which case mutation3D will attempt to find clusters across all proteins and models for which this mutation has an effect on protein products.

Users may elect to set the CL-distance, or the maximum allowable distance between α-carbons in a cluster of substituted amino acids. We refer to this as the *maximum cluster diameter* as this is equivalent to the maximum allowable diameter in Angstroms of a sphere encapsulating all α-carbons in a cluster. With regard to the complete linkage clustering algorithm, the CL-distance is the maximal dissimilarity between elements, after which, no new merging of elements and groups of elements occurs. In mutation3D, we call this parameter the *Maximum Clustering Diameter*, which is measured in Angstroms, and represents the maximum distance between amino acid substitutions after which no further merging of single mutations with clusters occurs and clusters are assigned based on current hierarchical groupings of mutations. For more information on all algorithm parameters and their default values, see Supp. Sections A.2 and A.3.

5.3.2 Statistical significance of clusters

In order to calculate the statistical significance of clusters found by complete-linkage clustering, mutation3D performs an iterative bootstrapping method to calculate a background distribution of cluster sizes arising from a random placement of an equivalent number of substitutions in a given protein structure. By default, mutation3D will randomly rearrange all amino acid substitutions 15,000 times in a given structure and calculate the minimum CL-distance at which a cluster of size *n* (where *n* is all cluster sizes found in the original data) is observed in the randomized data. For each cluster in the original data, *P*-values are computed empirically as the percentile rank of its CL-distance among all CL-distances for randomized clusters containing the same number of amino acid substitutions. The clustering algorithm/statistical significance calculator is implemented in C++ and is available for download as a command-line tool.

There is precedent, even within cancer gene detection, for the use of iterative bootstrapping methods when the background distributions are unclear or complicated (Hodis et al., 2012; Lawrence et al., 2014). Here we use bootstrapping to account for vastly different configurations of the protein backbone in different protein structures.

5.3.3 Compiling a protein structure and model set

In order to build a repository of protein structures and models, we curated experimentally-determined crystal structures from the PDB and homology models from ModBase by searching for canonical isoforms of Swiss-Prot structures or chains in both. Since many PDB structures provide too little coverage of their target protein to be useful for clustering, we retained only those structures that cover at least 250 amino acids or 40% of their target protein. We only retained ModBase models that have an MPQS score ≥ 0.5, and maintain a default cutoff of MPQS ≥ 1.1 in the mutation3D interface and in our analyses. All structures and models were compared against each other to remove redundancies (i.e. a ModBase model that is of higher quality than and whose range of amino acids is entirely contained within a second ModBase model derived from the same PDB structure was considered not to add any novel structural information to our repository). Furthermore, the amino acid indices of all models and structures were realigned using SIFTS (Velankar et al., 2013) to match the amino acid indices of the Swiss-Prot protein they represent.

5.3.4 mutation3D web interface

To build the mutation3D web interface, we leveraged the power and flexibility of several well-known JavaScript packages, such as JQuery and Bootstrap, in addition to a package designed to draw static two-dimensional figures (KineticJS). The cornerstone of our display system is an entirely

JavaScript-based molecular viewer, GLmol, which allows users to view interactive 3D protein structures natively in modern web browsers supporting the new WebGL standard, without downloading any additional software. We have made modifications to these software packages to allow triggering of events by the user, such as highlighting mutations and mutation clusters simultaneously in the 3D and 2D representations of proteins.

To speed up web accession for both single and batch queries, mutation3D runs on a multi-core web server and the calculation of clusters is distributed among available computing cores using multithreaded CGI programs.

5.3.5 Compiling mutations and variants affecting aromatase

We compiled a list of all inherited missense mutations from the Human Gene Mutation Database (Stenson et al., 2014) (HGMD) that (i) occurred within the exons of the *CYP19A1* gene [MIM# 107910] encoding the protein aromatase and (ii) have been shown in the primary literature to cause aromatase deficiency [MIM# 613546] (Supp. Table A.1). We also compiled a set of all missense SNPs with total minor allele frequency (MAF) ≥ 1% (combined African and European ancestry) from the Exome Sequencing Project (Fu et al., 2013) (ESP) that give rise to amino acid substitutions in aromatase (Supp. Table A.2). Please note that nucleotides are indexed in coding sequences, using the A of the ATG translation initiation start site as nucleotide 1. Visual inspection was performed by highlighting C_α positions in aromatase (PDB: 3S79) using PyMol (Schrodinger, 2010).

5.3.6 Segregating disease mutations from SNPs

For each Swiss-Prot protein from UniProt, a set of pathogenic inherited mutations from HGMD (Stenson et al., 2014) was assembled for the catalogued disease with the greatest number of

associated mutations in that protein. Proteins with fewer than three pathogenic mutations (two of which were required to occur at unique amino acid positions) associated with any one disease were not considered as this is the minimum requirement for identifying a cluster with default mutation3D parameters (Supp. Sections A.2 and A.3). Separately, we assembled non-synonymous SNPs (nsSNPs) with MAF ≥ 1% from the ESP 6500 set, only retaining proteins if there were at least three SNPs in the protein, two of which caused amino acid substitutions at unique amino acid positions. We intersected these two sets and only retained proteins that occurred in both sets as meeting the individual criteria of three mutations from each set, two of which must have been at unique amino acid positions, for a total of six or more variants per protein. In total, we retained 8,869 inherited disease-associated mutations from HGMD and 2,004 nsSNPs from ESP 6500 in 336 proteins.

We used mutation3D to identify clusters in the resulting proteins, employing a fairly strict definition of a cluster whereby a cluster was identified if three or more substitutions were found within the complete linkage clustering distance of 15 Å, with at least two substitutions occurring at unique amino acid locations. 3D model sets were derived from PDB structures and ModBase models indicated to be of high quality by an MPQS ≥ 1.1 (full details on default parameters for mutation3D are available in Supp. Sections A.2 and A.3). We report the average per-protein clustering rates across all proteins for which models from the correct set were available. *P*-values were calculated using a *U* test.

5.3.7 Measuring the overlap between mutation3D-implicated genes and the Cancer Gene Census

To assess whether mutation3D is able to report known cancer genes, we ran mutation3D with default parameters (Supp. Sections A.2 and A.3) on all WGS screens in COSMIC v75 (285 studies).

We varied the maximum cluster diameter from 5 Å to 25 Å and identified the fraction of proteins implicated (as having one or more clusters of amino acid substitutions) that are known cancer genes. We define known cancer genes to be the union of genes included in the Cancer Gene Census (Futreal et al., 2004) and MutSig drivers list (Lawrence et al., 2014). Overlaps between mutation3D-identified genes and known cancer genes were computed as the number of known cancer genes identified by mutation3D divided by the total number of genes implicated by mutation3D in each tissue category and overall (this is also known as the precision or positive predictive value (PPV)):

$$PPV = TP / (TP + FP)$$

where TP is the number of true positives and FP is the number of false positives predicted by mutation3D. It should be noted that since the our set of known cancer genes is far from complete, this estimation is likely to represent the lower bound of the true precision of our method. Furthermore, we acknowledge that even genes in the set of known cancer genes may not be drivers in all cancer types. However, the overlap between our results and the known cancer genes is likely to correlate with the underlying precision of our method and there is no reason to believe that the overlap will be biased in certain cancer types. Therefore, this measurement can be used to estimate the lower bound of the precision of our method in comparing its performance across different cancer types. Calculation of sensitivity and specificity is inappropriate in this instance because no method could re-capitulate all known cancer genes as no data set (single WGS study or a group of WGS) can be assumed to harbor all mechanisms underlying tumorigenesis. We also computed the overlap of *all* genes in these 285 COSMIC studies with known cancer genes for each tissue category and across all tissues, to show that performing 3D clustering at any maximum cluster diameter

increases precision over random expectation for this data set. *P*-values were calculated using a *Z* test to compare each fraction of identified genes by clustering at different diameter thresholds to the fraction of identified genes without clustering.

5.3.8 Assessing the likelihood of mutations clustered with mutation3D to be causal

In addition to predicting driver genes based on those found to contain clusters, mutation3D has the ability to predict those mutations likely to drive cancer phenotypes by their inclusion in clusters. Here, we used two proxies for causal driver mutations: that they should be more likely to be damaging and they should be more frequently observed in WGS studies.

We determined PolyPhen-2 scores (using the HumVar-trained model for assigning categories) of those mutations likely to be most deleterious biochemically based on a Grantham score (Grantham, 1974) in the top 25%. This shows how a combined biochemical and evolutionary genetics approach could lead to the discovery of new driver mutations. PolyPhen-2 scores were accessed using the Ensembl Variant Effect Predictor, assembly GRCh38.p5 (http://www.ensembl.org/Tools/VEP) (McLaren et al., 2010).

We further determined the fraction of mutations from WGS studies found in clusters that are observed at high frequencies (in the top 2%) throughout COSMIC WGS studies.

5.4 RESULTS

5.4.1 Single-protein spatial mutation case studies

The specific relationship between 3D regions of protein structure and their functions can be illustrated by the proximity of amino acid substitutions arising from known disease-causing and cancer-associated mutations in tertiary protein structures. We searched the Human Gene Mutation

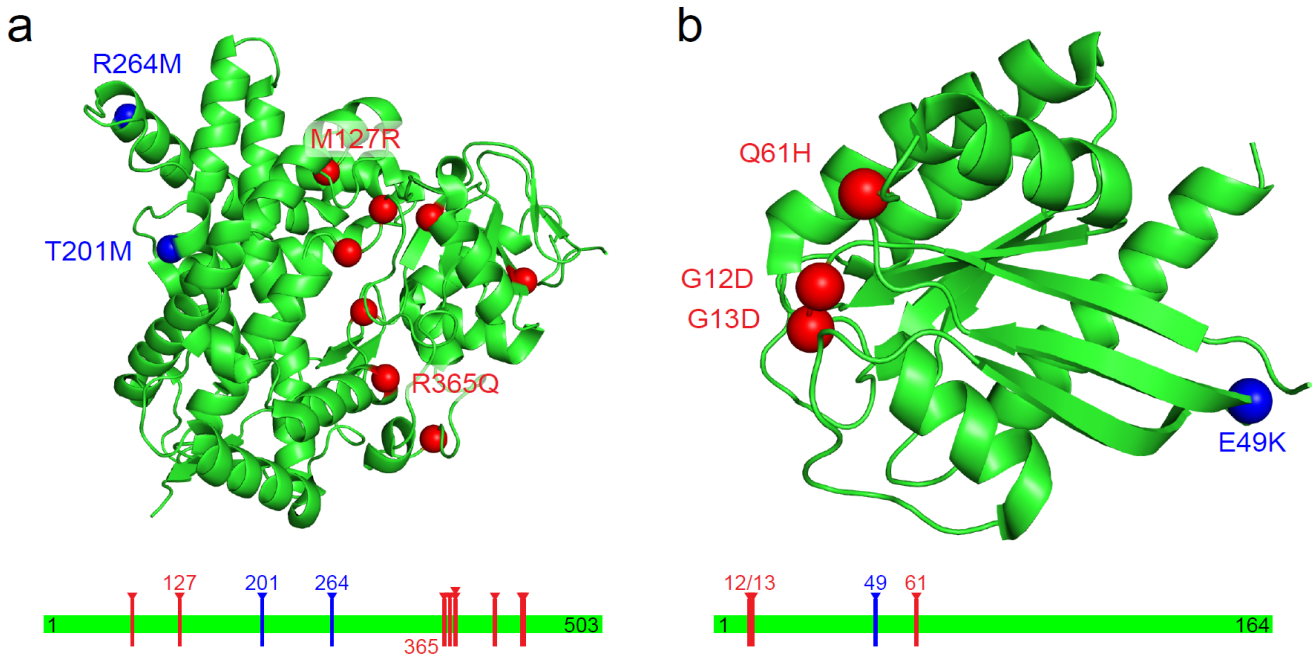

Figure 5.1 Missense mutations in both Mendelian disorders and cancer form clusters in tertiary protein structures. Linear protein models are given below each structure to illustrate the importance of studying 3D crystal structures. (a) Protein substitutions arising from mutations known to cause aromatase deficiency (in red) are shown overlaid upon an experimentally determined crystal structure of aromatase. Protein substitutions arising from common missense SNPs with MAF ≥ 1% (in blue) are shown to aggregate within regions of the protein structure distinct and spatially separated from those harboring the pathogenic substitutions, suggesting a strong relationship between the position of a substitution in the protein and its functional consequence(s). (b) Mutations causing amino acid substitutions in codons 12, 13 and 61 account for over 99% of the mutations in GTPase KRas reported by COSMIC. The three most common amino acid substitutions at these positions (shown in red) form a tight cluster in a crystal structure of GTPase KRas, whereas a substitution (E49K) only observed once in COSMIC (shown in blue), is likely to be a passenger mutation and falls outside the 3D mutation cluster even though it appears to be in close proximity in the linear model.

Database (Stenson et al., 2014) (HGMD), a large-scale disease database of gene mutations causing human inherited disease, and the Catalogue of Somatic Mutations in Cancer (Forbes et al., 2011) (COSMIC), a somatic cancer mutation database, for examples of spatially specific disruptions that might explain disease phenotypes. This is intended as a proof-of-principle, showing that there is a plausible connection between the spatial arrangement of mutations and disruptions of function, and that this relationship can be quickly captured through visual inspection.

Disease mutations and nsSNPs segregate in aromatase

According to HGMD, aromatase deficiency is known to be caused by at least 9 unique missense mutations in the cytochrome P450, family 19, subfamily A, polypeptide 1 (*CYP19A1*) gene leading

to amino acid substitutions at 8 positions along the aromatase protein backbone (Supp. Table A.1). The Exome Sequencing Project (ESP) 6500 data set (Fu et al., 2013) contains two common non-synonymous SNPs (nsSNPs) with MAF ≥ 1% in this gene, which we consider likely to be benign given their high frequency of occurrence (Supp. Table A.2). Based on the primary sequence alone, no clear pattern or separation can be detected between the disease mutations and nsSNPs (Figure 5.1a). However, when we inspect the locations of these two classes of mutation on an experimentally-determined crystal structure of aromatase (PDB: 3S79 in Figure 5.1a), it is evident that the verified disease mutations and common nsSNPs are localized in quite different regions of the protein, suggesting somewhat different functional consequences depending upon the location of a mutation within the tertiary structure of the protein.

Commonly observed cancer mutations form a tight cluster in GTPase KRas

Cancer mutations may also aggregate within clusters in protein structures, and this aggregation is likely to have profound implications for our ability to differentiate functional driver mutations from neutral passenger mutations. Consider the canonical oncogenic protein GTPase KRas: the tight clustering of commonly mutated amino acid substitutions in codons 12, 13 and 61 suggests that these mutations cause similar structural perturbations that may lead to many types of cancer (Figure 5.1b). In fact, it has long been known that substitutions in these codons confer tumorigenesis, and several mechanisms have been proposed (Pylayeva-Gupta et al., 2011) (Supp. Section A.4, Supp. Table A.3). Interestingly, another amino acid substitution E49K has only been reported once in a single patient (Guedes et al., 2013) and is predicted to be benign by PolyPhen-2 (Adzhubei et al., 2010). The clear spatial separation of the known driver mutations from the putatively benign mutation indicates a highly specific correlation between protein structure and

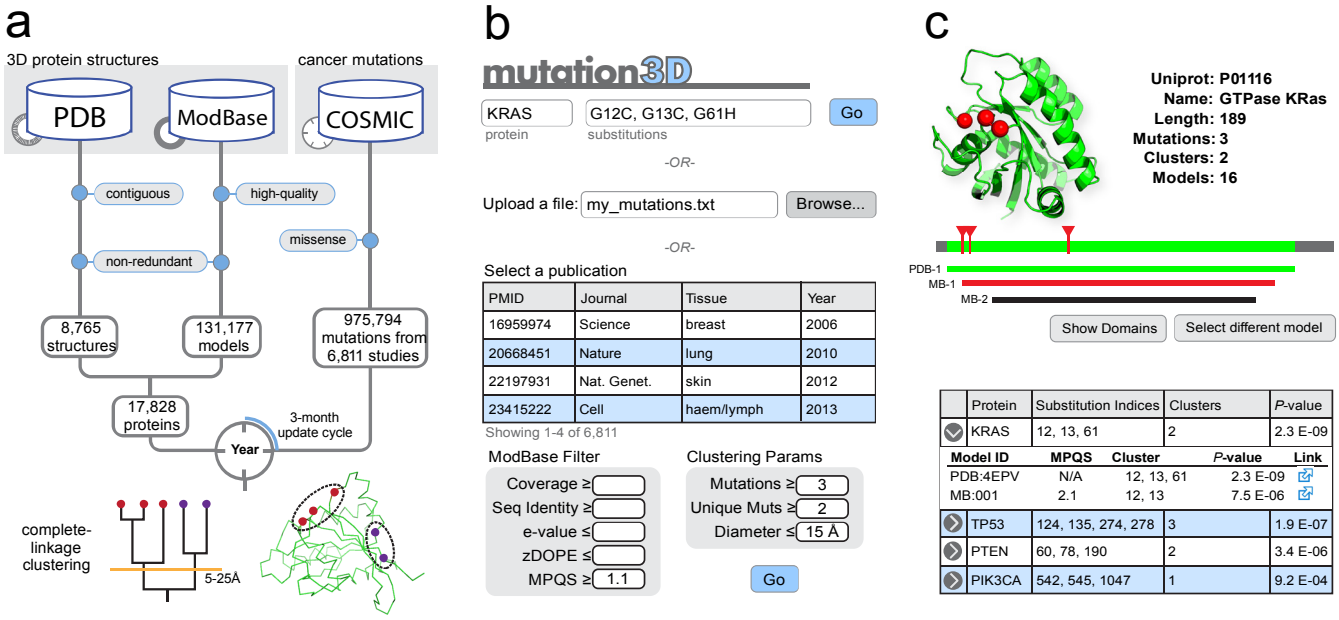

Figure 5.2 An overview of the mutation3D clustering and web accession procedures. (a) Sources of 3D protein structures and models and missense mutations in cancer. Pre-computation of clusters of amino acid substitutions for large data sets occurs with each COSMIC update. (b) There are three options for users to determine clusterings: by inputting their own data as substitutions in single proteins (or nucleotide mutations in genes), by uploading a file of mutations, or by analyzing missense mutations from one of the 6,811 publications curated by COSMIC. (c) The mutation3D web interface shows clusters on both linear models and interactive 3D models. Users may select among available models and structures. Individual queries will lead directly to this page, while batch queries will first lead to a table of proteins and clusters (shown below).

function in cancer. Owing to its very high mutation frequency in many different types of cancer, *KRAS* [MIM# 190070] is readily identifiable as tumorigenic by many methods; however, mutation3D is uniquely positioned to be able to detect similar cases of spatially specific disruption in proteins currently unknown for their roles in tumorigenesis by relating cancer sequencing data to aberrations in the structural proteome.

5.4.2 Coordinating mutations and structural data into a tool for whole-genome inference

mutation3D identifies mutations that group together to form statistically significant clusters on the folded protein backbone based on atomic coordinates derived from experimentally determined crystal structures and homology models. Cluster significance is measured by an iterative bootstrapping model, in which observed mutations are randomly rearranged on a protein structure,

and the size of the observed cluster is ranked compared to all randomly derived clusters to compute an empirical *P*-value (see Methods for details). The accompanying web interface provides visualization of these clusters as well as the ability to rapidly switch views between all available structures. Figure 5.2 describes the curation of structural and mutation data, and user accession and download procedures.

Structural data underlying mutation3D

In assembling a set of protein structures and models for use with mutation3D, we relied on the huge advances made in structural proteomics over the past decade. Alongside the explosion of genomic sequencing data, the availability of structural proteomic data, including crystal structures and homology models, has increased dramatically. In 2003, there were 25,864 crystal structures in the Protein Data Bank (Berman, 2000) (PDB), covering 6.7% of the human proteome. Now, with the number of entries in the PDB exceeding 100,000, we can visualize nearly 90% (with reasonable accuracy and coverage—see Supp. Figure A.1) of the human proteome through a combination of experimentally-determined crystal structures and structural models based on shared structural elements among homologous proteins. mutation3D curates both crystal structures from the PDB and high-quality homology models from ModBase (Pieper et al., 2011) to populate its repository of over 135,000 protein structures (Figure 5.2a). This significant underpinning of structural proteomic data ensures that mutation3D is useful for large-scale sequencing projects, as nearly all DNA mutations of interest within coding regions will be mappable to 3D locations in protein structures.

Seamless access to large-scale somatic cancer mutation sets

Perhaps the richest large-scale source of missense mutation data derives from WGS studies of cancer patient cohorts. According to COSMIC, in the year 2003, 187 peer-reviewed articles were published reporting on average a single gene with protein-altering somatic mutations in tumor-normal sequencing studies. In 2012, 572 studies reporting an average of 144 mutated genes were published. With the growing ease of sequencing, the scientific community has largely embraced the wholesale sequencing of tumor samples, and an accompanying class of statistical methods to identify genes characterized by elevated mutation rates across large patient cohorts (Cancer Genome Atlas, 2012; Hodis et al., 2012; Lawrence et al., 2014; Lawrence et al., 2013; Sjöblom et al., 2006; Wood et al., 2007). These methods have been largely successful, and have led to the discovery of many genes previously not known to be involved in tumorigenesis. However, studying cancer at the level of whole genes ignores the fact that many genes and their protein products perform multiple cellular functions (pleiotropy). By incorporating available protein structures and models into cancer gene detection, we can harness the inherent structure-function relationship in proteins to identify more specific tumorigenic etiologies based on specific spatial disruptions that could become therapeutic targets.

The mutation3D web interface allows users to rapidly analyze pre-processed missense mutation data from the most recent build of COSMIC through intuitive web forms on the *Advanced* query page (http://mutation3d.org/advanced_form.shtml, click the *COSMIC* tab under *Data Source*). Currently, we have catalogued over 975,000 missense mutations in 6,811 primary cancer sequencing studies that users can search for by author, journal, PMID, and size of dataset (Figure 5.2b). Additionally, users may choose to tune the default clustering parameters (Supp. Section A.3) and protein structural model set (Supp. Section A.2) based on the types of evidence needed to support clusters

for their specific application. A list of candidates, with links to 3D views of the mutations overlaid onto structural models (Figure 5.2c), are retrieved within seconds, even for the largest WGS studies in COSMIC.

5.4.3 mutation3D identifies well-validated gene candidates and plausible new targets

We ran mutation3D on large sets of known inherited disease and cancer mutations to demonstrate the power of clustering to reveal shared etiologies in the structural proteome. Here, and in all following large-scale analyses, mutations associated with each distinct disease phenotype are considered separately from mutations associated with unrelated phenotypes so that a correspondence can be made between clusters in functionally relevant parts of protein structures and potential defects in molecular function that may cause one specific disease or type of cancer. We demonstrate the ability of mutation3D to distinguish functional from non-functional mutations in disease and to re-discover many known cancer-causing genes as well as discovering several new putative targets. Parameters for all tests performed are available in Methods and in Supp. Table A.4.

mutation3D distinguishes disease mutations from common variants

To illustrate the efficacy of mutation3D in distinguishing functional from non-functional variants, we considered all proteins harboring at least 3 mutations associated with a single disease (according to HGMD) and all missense population variants (SNPs) from the ESP 6500 data set for this same set of proteins (see Methods for details). We were able to show that the resulting set of 8,869 disease-causing amino acid substitutions are more likely to be clustered by mutation3D than are 2,004 putatively benign substitutions arising from missense SNPs when considering only those mutations associated with a single disease at a time mixed together with SNPs in the same

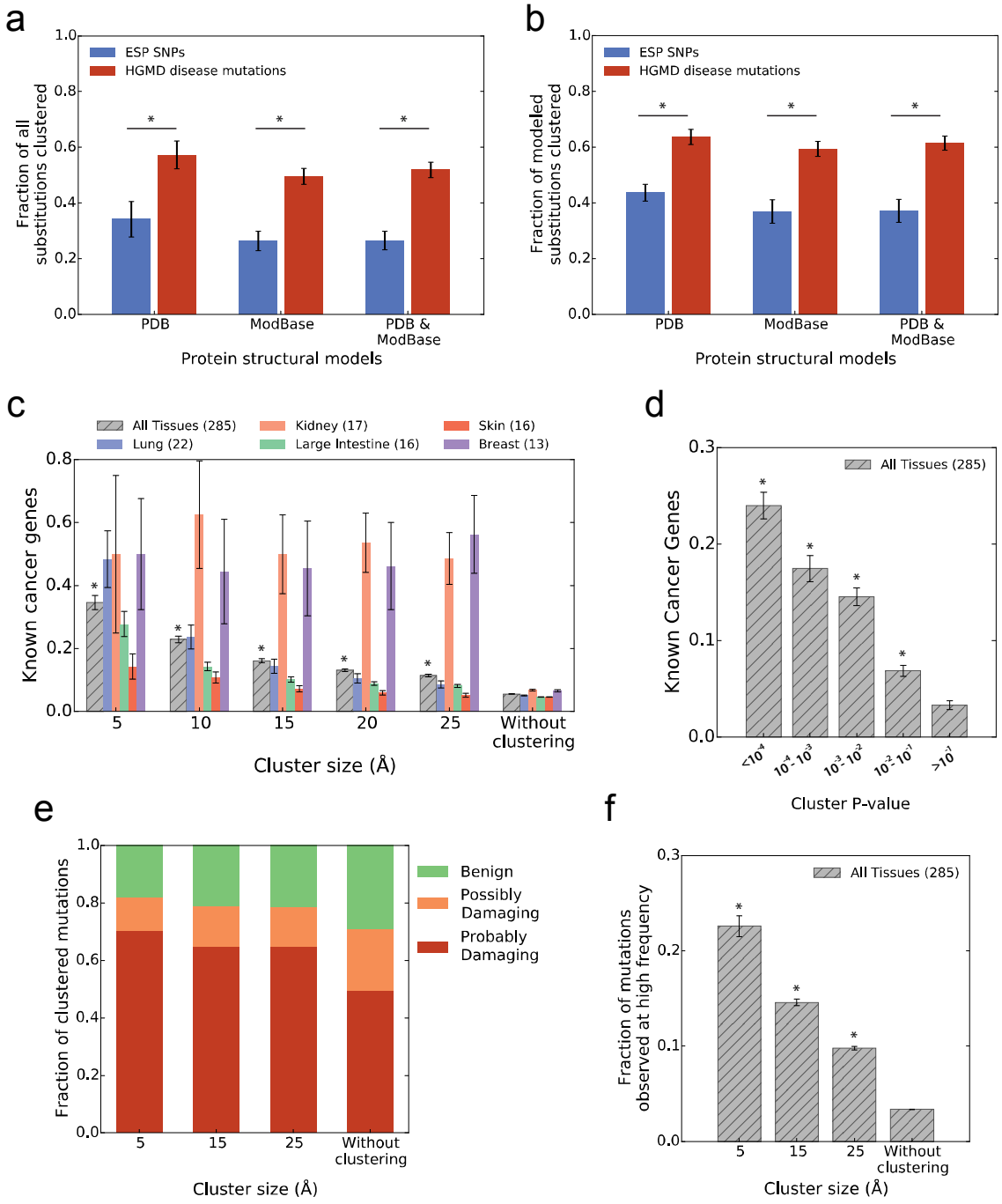

Figure 5.3 (a-b) Known inherited disease-associated missense mutations from HGMD and missense SNPs from ESP 6500 with MAF ≥ 1% were clustered using mutation3D, with the percentage of variants within proteins containing clusters reported. (a) The combined set of resulting amino acid substitutions was plotted onto 3D protein models derived from the PDB alone, ModBase alone, and a combination of the two. (b) Fractions of clustered mutations were recalculated only for those mutations that reside within protein regions for which a 3D structure or model exists. (c-e) mutation3D was run on 285 WGS somatic tissue screens in COSMIC. (c) A higher fraction of protein candidates identified are known cancer genes at smaller values of cluster size (maximum cluster diameter). (d) A higher fraction of protein candidates identified are known cancer genes at smaller clustering P-values. (e) Mutations in tighter clusters are predicted by PolyPhen-2 to be more damaging than those in sparser clusters and in all WGS studies. (f) Mutations in tighter clusters are more likely to be observed at high frequency across COSMIC WGS studies. For all panels, * indicates P < 0.01 comparing categories marked with bars (in a and b), or to the final category of the plot ("Without clustering" in c and f; " >10-1" in d).

proteins (Figure 5.3a-b). This trend is apparent irrespective of whether the protein structure set is confined to known PDB structures, homology models from ModBase, or a combination of the two.

This analysis illustrates mutation3D's ability to distinguish functional from non-functional variants when all functional variants share an associated phenotypic consequence. Because it is often difficult to determine which cancer mutations are drivers and which are passengers, mutation3D's ability to distinguish functional disease mutations from non-functional SNPs serves as a proxy measure of its ability to separate functional driver mutations from a background of largely non-functional passenger mutations.

mutation3D identifies both new and well-known cancer genes

To confirm that mutation3D identifies plausible driver gene candidates in cancer (as judged by the existence of one or more clusters of substitutions in structures of their protein products), we computed statistically significant clusters from mutations in all WGS studies cataloged by COSMIC. First, we calculated the proportion of the identified cancer candidates that have been previously proposed as cancer drivers based on a combination of the Cancer Gene Census database (Futreal et al., 2004) and the MutSig driver list (Lawrence et al., 2014). This is likely to be correlated with the lower bound of precision, or positive predictive value, of our method (see Methods). Figure 5.3c illustrates the calculated proportion values for all publications analyzed and for specific tissues within these studies, plotted over several cluster sizes. The results concur with our expectation that tighter mutation clusters should exhibit high precision for known cancer genes since substitutions in close physical proximity will be more likely than distant substitutions to be contained within the same interface domain or within the hydrophobic protein core. As expected, we also observe lower precision in the identification of genes involved in cancers of the skin, which are characterized by

very high mutation rates (Alexandrov et al., 2013). By contrast, cancers of the breast are known to harbor driver mutations in a relatively small number of genes and contain a relatively low proportion of passenger mutations (Kan et al., 2010), thereby allowing mutation3D to precisely identify known cancer genes irrespective of cluster size.

To confirm that our statistical model yields plausible measures of cluster significance, we computed the statistical significance of clusters found in COSMIC WGS data. We find that our iterative bootstrapping model (See Methods) produces *P*-values that are highly correlated with the likelihood of a gene to be a known cancer genes (Figure 5.3d).

We also find that the somatic mutations within these clusters are predicted to be more deleterious by PolyPhen-2 when found in smaller, more specific clusters (Figure 5.3e). Furthermore, mutations within clusters are observed at much higher frequencies within WGS studies, suggesting they are likely to be driver mutations (Figure 5.3f). Overall, these analyses suggest a tendency for functionally important mutations to form clusters in cancer patient cohorts, whereas less important passenger mutations are more likely to fall outside these clusters.

We next investigated whether mutation3D preferentially reports potential oncogenes or tumor suppressors. We find that of genes annotated in either class based on the Cancer Gene Census, there is not a significant difference in the likelihood mutation3D will find clusters within their protein products (Supp. Section A.5, Supp. Figure A.2). This suggests that mutation3D is equally robust in its ability to detect oncogenes and tumor suppressors.

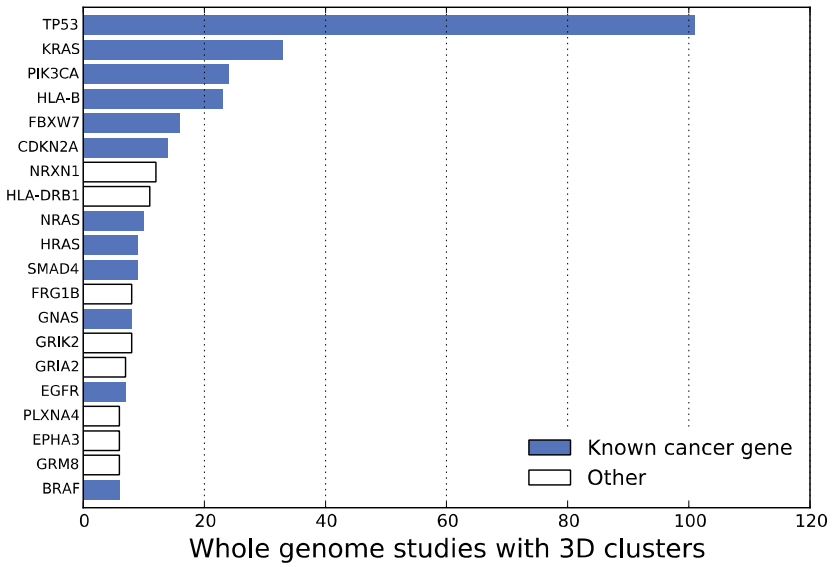

Figure 5.4 The top 20 genes implicated in WGS studies by mutation3D ranked by the number of publications in which clusters were observed for each.

Finally, we produced a list of the genes whose protein products most commonly exhibit clusters of mutations within the same set of COSMIC WGS publications. We find that mutation3D implicates many well-known cancer genes (*TP53, KRAS, EGFR, BRAF*, etc.) as well as some genes that are missing from the Cancer Gene Census (Figure 5.4). Visual inspection of the most significant clusters for each of these proposed genes demonstrates the power of 3D clustering (Figure 5.5). A list of all genes found in at least 4 studies across COSMIC is available in Supp. Table A.5.

5.5 Discussion

Researchers have already begun to acknowledge the added benefit of linear clustering approaches to the detection of driver mutations in two recently proposed methods (Lawrence et al., 2014; Tamborero et al., 2013). However, these methods do not take into account the 3D positions of mutations within protein products, disregarding information available due to structure-function relationships in proteins. Two other recent methods (Ryslik et al., 2013, 2014) perform 1D

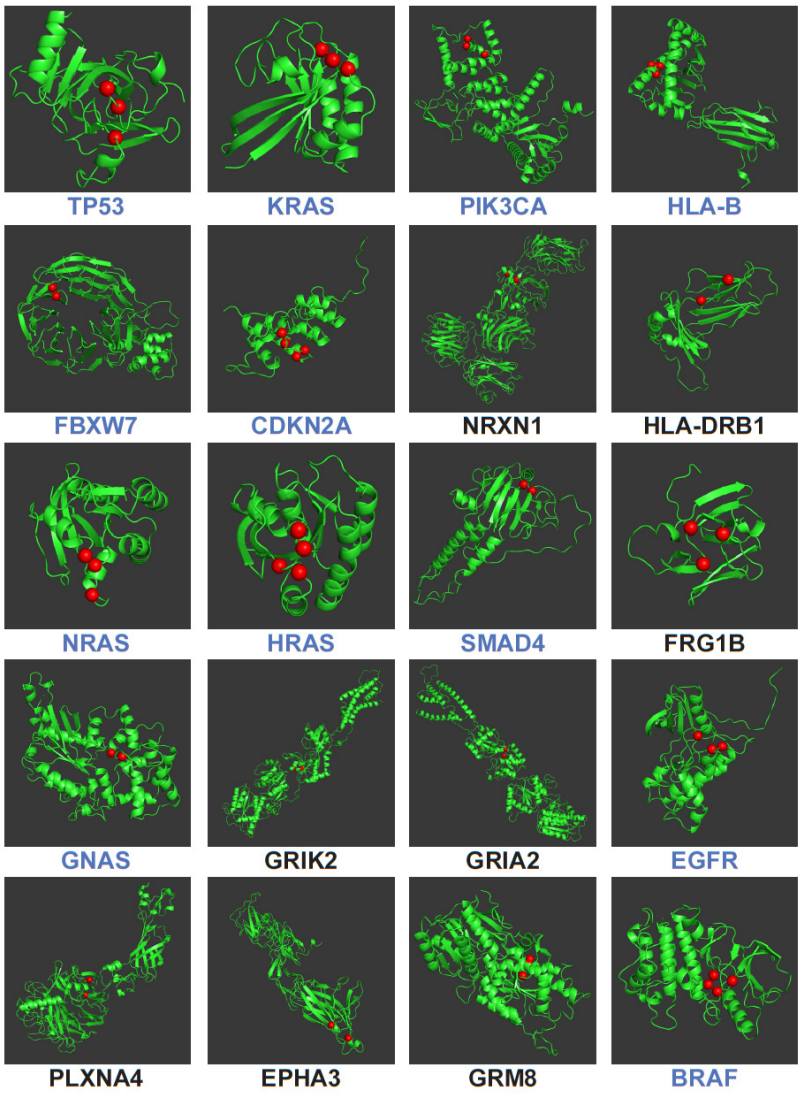

Figure 5.5 The most significant cluster for each of the top 20 implicated genes is shown in 3D.

clustering of mutations after a projection of 3D structural coordinates into 1D, potentially resulting in loss of information (Supp. Section A.6, Supp. Figure A.3). Clustering methods have also been used to detect signatures of positive selection (Tusche et al., 2012; Wagner, 2007; Zhou et al., 2008); however, the goals and assumptions of these methods and mutation3D are quite different (Supp. Section A.7, Supp. Figure A.4). Another recent method detects non-random distributions of mutations in protein crystal structures (Kamburov et al., 2015). Although this method shows in

principle that 3D structural information is valuable for identifying target genes, it does not distinguish individual clusters and its analysis is limited to PDB structures.

Compared to the standard class of methods that do not search for clusters of amino acid substitutions, but instead employ measures of mutation frequency at the gene level to detect drivers of cancer, the added value of mutation3D lies in its orthogonal use of protein structures to make a more direct connection between alterations of structure and disruptions of function. We do not intend that mutation3D should replace these methods (Hodis et al., 2012; Lawrence et al., 2013; Sjöblom et al., 2006). Instead, mutation3D gives scientists the ability to inspect their data through an additional lens—to visualize and form hypotheses about functional gene and protein candidates proposed by any method of cancer gene detection, and to find cases in which directly searching for structural disruptions may provide insights not available by other means. Even beyond its potential to improve candidate gene identification, mutation3D is valuable simply in terms of its ability to display mutations on all available high-quality structures and models, a task that requires significant effort on any scale without mutation3D, and can be accomplished on massive scales with mutation3D.

Throughout this study, we have evaluated the ability of mutation3D to identify *whether or not* a gene is involved in cancer because this is a standard for the cancer gene detection methods of today. However, such a metric may underrate the true ability of mutation3D, which can propose specific tumorigenic etiologies based on the structural localization of mutations. Even in cases where mutation3D identifies the same gene as another method, analyzing and viewing the mutations using mutation3D may present a specific hypothesis supported by both statistical and structural evidence, which may be more likely to inspire follow-up studies.

In addition to providing structural evidence for single proteins, the mutation3D web interface (http://mutation3d.org) allows users to rapidly search for clusters of mutations in the proteome (by inputting their data in a variety of popular genomic and proteomic formats), view, and download clustering reports. Through the Advanced Query interface, users may adjust the clustering parameters and build structure and model sets for custom analysis of their own data or to seamlessly access pre-analysis of over 975,000 missense mutations in 6,811 primary cancer studies catalogued by COSMIC. Owing to the amount of data already available via the mutation3D web interface and the continual accumulation of cancer sequencing and protein structural data, mutation3D is likely to produce future insights based on structural localization of mutations in the human proteome.

5.6 References

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Kondrashov, A., and Sunyaev, S. (2010). A method and server for predicting damaging missense mutations. Nature Methods *7*, 248-249.

Alexandrov, L., Nik-Zainal, S., Wedge, D., Aparicio, S., Behjati, S., Biankin, A., Bignell, G., Bolli, N., Borg, A., Børresen-Dale, A.-L.*, et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

Berman, H.M. (2000). The Protein Data Bank. Nucleic Acids Research *28*.

Cancer Genome Atlas, N. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.

Das, J., Fragoza, R., Lee, H.R., Cordero, N.A., Guo, Y., Meyer, M.J., Vo, T.V., Wang, X., and Yu, H. (2014a). Exploring mechanisms of human disease through structurally resolved protein interactome networks. Mol Biosyst *10*, 9-17.

Das, J., Lee, H.R., Sagar, A., Fragoza, R., Liang, J., Wei, X., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014b). Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. Hum Mutat *35*, 585-593.

Forbes, S., Bindal, N., Bamford, S., Cole, C., Kok, C., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research *39*, 50.

Fu, W., O'Connor, T., Jun, G., Kang, H., Abecasis, G., Leal, S., Gabriel, S., Rieder, M., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Futreal, P., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. (2004). A census of human cancer genes. Nature Reviews Cancer *4*, 177-183.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science *185*, 862-864.

Guedes, J., Veiga, I., Rocha, P., Pinto, P., Pinto, C., Pinheiro, M., Peixoto, A., Fragoso, M., Raimundo, A., Ferreira, P.*, et al.* (2013). High resolution melting analysis of KRAS, BRAF and PIK3CA in KRAS exon 2 wild-type metastatic colorectal cancer. BMC Cancer *13*, 169.

Hanahan, D., and Weinberg, R. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Hodis, E., Watson, I., Kryukov, G., Arold, S., Imielinski, M., Theurillat, J.-P., Nickerson, E., Auclair, D., Li, L., Place, C.*, et al.* (2012). A landscape of driver mutations in melanoma. Cell *150*, 251-263.

Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proceedings of the National Academy of Sciences of the United States of America *112*, E5486-5495.

Kan, Z., Jaiswal, B., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H., Yue, P., Haverty, P., Bourgon, R., Zheng, J.*, et al.* (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. Nature *466*, 869-873.

Kucukkal, T.G., Petukh, M., Li, L., and Alexov, E. (2015). Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Current opinion in structural biology *32*, 18-24.

Lawrence, M., Stojanov, P., Mermel, C., Robinson, J., Garraway, L., Golub, T., Meyerson, M., Gabriel, S., Lander, E., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.

Lawrence, M., Stojanov, P., Polak, P., Kryukov, G., Cibulskis, K., Sivachenko, A., Carter, S., Stewart, C., Mermel, C., Roberts, S.*, et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069-2070.

Miller, Martin L., Reznik, E., Gauthier, Nicholas P., Aksoy, Bülent A., Korkut, A., Gao, J., Ciriello, G., Schultz, N., and Sander, C. (2015). Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. Cell Systems *1*, 197-209.

Muller, P., and Vousden, K. (2013). p53 mutations in cancer. Nature Cell Biology *15*, 2-8.

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B.A., Hashimoto, K., Alexov, E., Wuchty, S., and Panchenko, A.R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. PloS one *8*, e66273.

Petukh, M., Kucukkal, T.G., and Alexov, E. (2015). On human disease-causing amino acid variants: statistical study of sequence and structural patterns. Hum Mutat *36*, 524-534.

Pieper, U., Webb, B., Barkan, D., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E., Pettersen, E., Huang, C.*, et al.* (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research *39*, 74.

Pylayeva-Gupta, Y., Grabocka, E., and Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. Nature Reviews Cancer *11*, 761-774.

Ryslik, G.A., Cheng, Y., Cheung, K.-H.H., Modis, Y., and Zhao, H. (2013). Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics *14*, 190.

Ryslik, G.A., Cheng, Y., Cheung, K.-H.H., Modis, Y., and Zhao, H. (2014). A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics *15*, 86.

Schrodinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Sjöblom, T., Jones, S., Wood, L., Parsons, D., Lin, J., Barber, T., Mandelker, D., Leary, R., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.

Sneath, P. (1957). The application of computers to taxonomy. Journal of General Microbiology *17*, 201-226.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr *5*, 1-34.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet *133*, 1-9.

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics *29*, 2238-2244.

Tusche, C., Steinbruck, L., and McHardy, A.C. (2012). Detecting patches of protein sites of influenza A viruses under positive selection. Mol Biol Evol *29*, 2063-2071.

Velankar, S., Dana, J., Jacobsen, J., van Ginkel, G., Gane, P., Luo, J., Oldfield, T., O'Donovan, C., Martin, M.-J., and Kleywegt, G. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research *41*, 9.

Vucic, E., Thu, K., Robison, K., Rybaczyk, L., Chari, R., Alvarez, C., and Lam, W. (2012). Translating cancer 'omics' to improved outcomes. Genome Research *22*, 188-195.

Wagner, A. (2007). Rapid detection of positive selection in genes and genomes through variation clusters. Genetics *176*, 2451-2463.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N., *et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.

Wood, L., Parsons, D., Jones, S., Lin, J., Sjöblom, T., Leary, R., Shen, D., Boca, S., Barber, T., Ptak, J., *et al.* (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108-1113.

Zhou, T., Enyeart, P.J., and Wilke, C.O. (2008). Detecting clusters of mutations. PloS one *3*, e3765.

CHAPTER 6

A pan-interactome map of protein interaction interfaces

6.1 ABSTRACT

Protein interactions underlie nearly all known cellular function, making knowledge of their binding conformations paramount to understanding the physical workings of the cell. However, experimentally determining atomic-level structures of interacting proteins in complex is expensive, and structures are only available for a small fraction of the known interactome. Furthermore, previous methods to predict interaction interfaces sacrifice either quality or scale, precluding genomic-scale analyses of interaction interfaces. Here we present ECLAIR, an ensemble machine learning algorithm to predict the interface of any protein interaction based on the most informative set of features available, including structural and partner-specific features. We used ECLAIR to predict interfaces for 118,113 protein interactions with previously unresolved interfaces in human and 7 model organisms, a 10-fold increase over previously known interfaces. We find that predicted interfaces share several functional properties of known interfaces, including an enrichment for disease mutations, suggesting their applicability for functional genomic studies. Through 2,164 mutagenesis experiments we show that mutations of predicted interface residues disrupt interactions at a similar rate to known interface residues and at a much higher rate than mutations outside of predicted interfaces. Finally, we combined our predictions with known structural interfaces to produce a pan-interactome resource containing the highest quality interfaces available for 130,634 interactions in 8 organisms (http://eclair.yulab.org).

6.2 INTRODUCTION

Protein-protein interactions facilitate much of known cellular function. Recent efforts to experimentally determine protein interactomes in human (Rolland et al., 2014) and model organisms (Arabidopsis Interactome Mapping, 2011; Vo et al., 2016; Yu et al., 2008), in addition to literature curation of small-scale interaction assays (Das and Yu, 2012), have dramatically increased the scale of known interactome networks. Studies of these interactomes have allowed researchers to elucidate how modes of evolution affect the functional fates of paralogs (Vo et al., 2016) and to examine on a genomic scale network interconnectivities that determine cellular functions and disease states (Sahni et al., 2015).

While simply knowing which proteins interact with each other provides valuable information to spur functional studies, far more specific hypotheses can be tested if the spatial contacts of interacting proteins are known (Kim et al., 2006). For instance, it has been shown that disease mutations tend to localize to interacting protein domains and mutations on the same protein may cause clinically-distinct diseases by disrupting interactions with different partners (Wang et al., 2012). However, the binding topologies of interacting proteins (i.e. the relative positions of all atoms in an interaction interface) can only be absolutely determined through resource-intensive X-ray crystallography, NMR, and more recently cryo-EM (Kuhlbrandt, 2014) experiments, severely limiting the number of interactions with resolved interaction interfaces.

Computational methods have been employed to predict the atomic-level bound conformations of interactions whose experimental structures have not yet been determined. For instance, molecular docking uses resource intensive physics simulations to find low energy conformations for individual structures of the two interacting proteins (Halperin et al., 2002). However, while it is capable of producing high quality interaction models (Lensink et al., 2016), docking remains highly

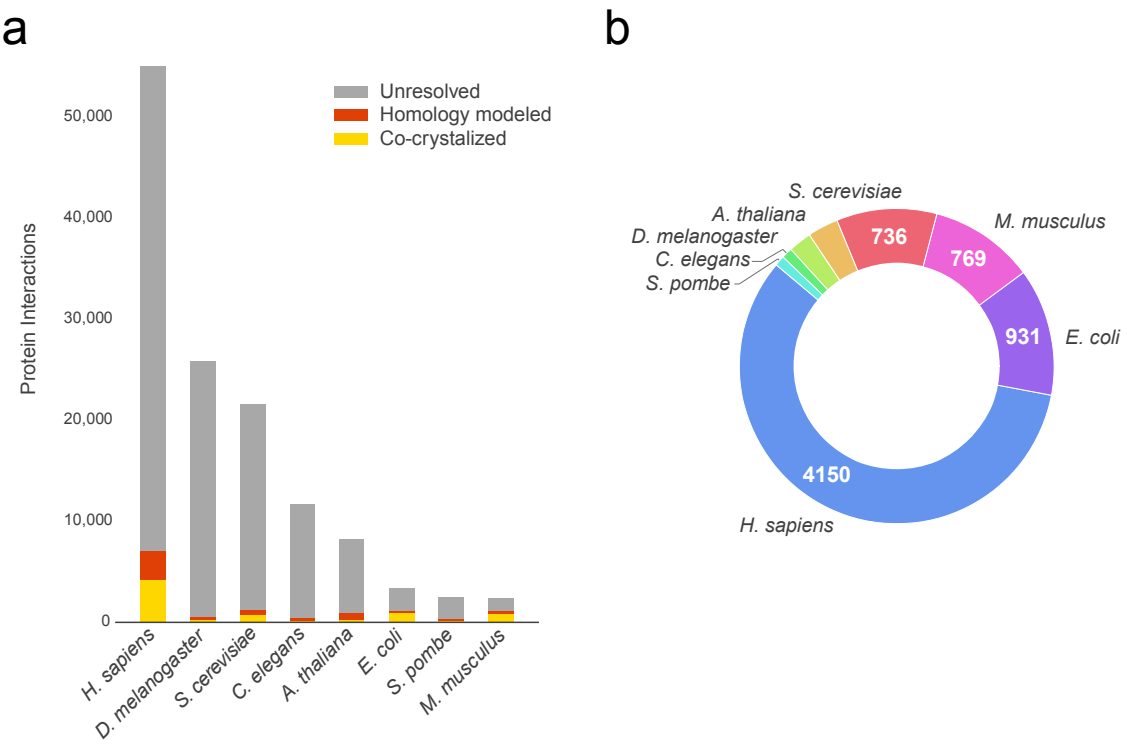

Figure 6.1 The current size of structural interactomes. (a) The sources of pre-computed structural interactomes and their coverage of known high quality binary interactomes. (b) Interactions from the largest 8 interactomes with experimentally solved structures.

specialized, and models are not yet available on a large scale. Homology modeling, on the other hand, requires aligning interacting proteins to a structural template, and has been used to produce models of all interactions in commonly studied organisms for which templates are available (Mosca et al., 2013). However, due to the requirement of structural templates, homology modeling can only be used for <5% of known interactions. Together, co-crystal structures and homology models comprise the currently available pre-calculated sources of structural interactomes (**Figure 6.1**).

While determining the atomic-resolution of protein interactions is an ultimate goal of structural biology, current methods have thus far provided relatively few structures and models of protein interactions compared to the size of known interactomes. For the remaining ~90% of interactions without structural information, a lower-resolution picture of interfaces can provide crucial information for functional studies, and help to complete structural interactome networks to the best of our current capabilities (Vakser, 2013). For instance, residue-level interaction interfaces, where

we know which residues are at the interface, but not their precise structural arrangement, can be a great boon to genomic-scale functional analyses (Brunk et al., 2016; Xie et al., 2014), and elucidate common modes of human disease (Das et al., 2014; Wang et al., 2012). Furthermore, an interactome network containing the highest possible resolution of protein interaction interfaces can be an indispensable tool for targeted studies to elucidate pathways and dissect disease mechanisms (Vo et al., 2016; Wei et al., 2014).

Here, we present ECLAIR, an **e**nsemble **c**lassifier **l**earning **a**lgorithm to predict **i**nterface **r**esidues of protein-protein interactions. Unlike other machine learning approaches that have been applied to predicting interface residues, ECLAIR is designed to be applicable to *any* interaction using all available features for each. ECLAIR is also the first prediction algorithm to have been used to precompute interaction interfaces for entire interactomes—we predict interface residues for all interactions in 8 highly studied interactomes (*H. sapiens*, *D. melanogaster*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, *E. coli*, *S. pombe*, and *M. musculus*). Furthermore, ECLAIR combines recently proven advances in co-evolution- and docking-based feature construction (Hopf et al., 2014; Hwang et al., 2014) for predicting interfaces of specific binding partners, as well as advances in machine learning algorithm optimization (Bergstra et al., 2011), which allowed us to tune each classifier in the ensemble independently to ensure maximum performance.

ECLAIR (http://eclair.yulab.org) is deployed as an interactive web server, containing predicted interface residues for 118,113 previously un-resolved interactions in human and 7 model organisms, a 10-fold increase over previously known interfaces. Furthermore, for 12,521 interactions with pre-existing sources of structural evidence (co-crystal structures or homology models), we calculate interface residues and display interactive 3D models. Together, interface residues from experimentally-determined structures, homology models, and our predicted

interfaces form a unified, full-coverage map of protein interfaces, giving scientists access to the highest quality interface evidence available for all interactions in these organisms, and for the first time make possible analyses of protein interaction interfaces on a genomic scale.

6.3 RESULTS

ECLAIR addresses a pervasive shortcoming in current methods to predict protein interaction interfaces that may preclude their use in genomic studies—that they sacrifice either interactome coverage or prediction quality in pursuit of the other. Methods to date either use information-rich features (i.e. structurally-derived features) to obtain the highest accuracy on a small fraction of interactomes, or they use widely available features in order to predict on a large scale, but at a severe cost to accuracy as they must ignore rich structural features, which are not available for entire interactomes (Esmaielbeiki et al., 2016; Ezkurdia et al., 2009; Zhou and Qin, 2007).

The source of this dichotomy in currently available methods for predicting protein interfaces is the sparsity of features available for building machine learning classifiers (Esmaielbeiki et al., 2016; Ezkurdia et al., 2009; Zhou and Qin, 2007). Since most classifiers are ill-equipped to handle even a single missing value in a feature array, methods to date have been largely focused on showing the value of new features or techniques on sets of interactions that satisfy the feature requirements of a classifier. As a result, the field of interface prediction has been well-explored, but there is still no unified dataset containing interfaces for entire interactomes, and indeed no method capable of predicting for all interactions without sacrificing prediction quality by ignoring sparse experimentally derived features.

To spur genomic studies of interaction interfaces, which require both high coverage and high quality, we propose a new method, ECLAIR, which is capable of classifying interfaces for all

interactions using the richest subset of features for each interaction. To accomplish this, we implement an ensemble of independently trained classifiers, each covering a common feature subspace, that together ensure all interactions can be classified using a single classifier trained on the most information-rich set of features.

6.3.1 Feature selection and engineering

In order to build a machine learning classifier to predict whether or not an amino acid residue is at the interface of an interaction, we selected an initial set of features widely reported (Esmaielbeiki et al., 2016) to be predictive of interface residues: (1) ensemble biophysical features, such as residue hydrophobicity and polarity, (2) evolutionary sequence conservation, and (3) solvent-accessible surface area (SASA) (**Figure 6.2a**). Of these features, SASA is widely reported to be most informative, since all interface residues, by definition, are on the surface of proteins. However, methods using SASA typically take experimentally-determined structures as inputs, severely limiting their applicability to full length proteins and to those without structural coverage in the PDB. To combat this issue and vastly increase structural coverage, we also compute SASA from single protein homology models (Pieper et al., 2011).

One key missing aspect from this set of features and from many interface predictors is the incorporation of interaction partner-specific information. Methods that rely on just the aforementioned features will predict the same interface residues for a protein, regardless of its binding partner. To our knowledge, there are no current interface prediction methods that combine partner-specific features with other protein features to predict on a large scale, but several recent studies have described suitable methods for calculation of partner-specific features (Hopf et al.,

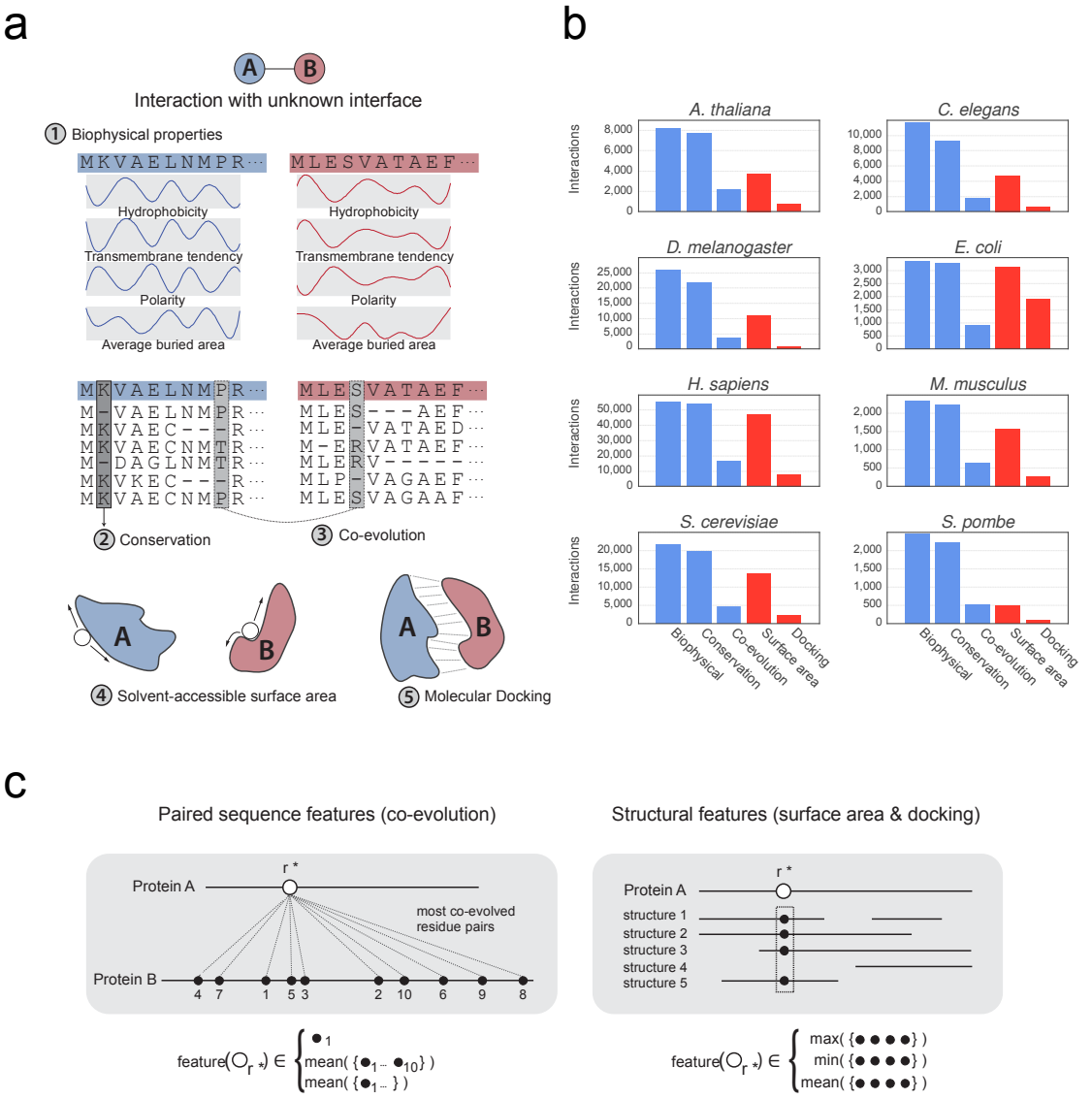

Figure 6.2 Features for predicting protein interaction interfaces. (a) A schematic showing the five feature categories from which feature sets are optimized to train ECLAIR. (b) The portions of interactomes for which each feature type is available. (c) Feature aggregation strategies employed for combining multiple points of evidence into single co-evolution- and structure-based features. For co-evolution, we select the top co-evolved residue, the mean of features for the top 10 co-evolved residues, or the mean over all co-evolved residues in the partner protein. For proteins with multiple structures, we take the mean, minimum, or maximum SASA over all available structures.

2014; Hwang et al., 2014). To build ECLAIR, we incorporate two classes of partner-specific features: (1) co-evolution of amino acid sequences and (2) computational molecular docking (**Figure 6.2a**).

The basis for use of evolutionary sequence-based features is the postulate that highly conserved sites in orthologous proteins are likely to be functionally important, and since much of protein function is determined by interactions with other proteins, conserved residues are more likely to

be interface residues (Guharoy and Chakrabarti, 2010). Co-evolutionary features summarize dependent patterns of conservation in two interacting proteins. In addition to being well-conserved, we expect that residues important to maintaining the interaction interface should evolve in a coordinated manner so as to maintain binding complementarity. Proven methods of measuring co-evolution include statistical coupling analysis (SCA) (Lockless and Ranganathan, 1999) and direct coupling analysis (DCA) (Morcos et al., 2014), both of which are used to produce features to train our classifiers.

Molecular docking is a class of stand-alone method to predict bound conformations of protein interactions by computationally sampling low-energy binding conformations of individual structures of interacting proteins. Docking is highly specialized and computationally intensive, making it ill-suited to interactome scale prediction. However, we incorporate docking results into our classifier by synthesizing features that attempt to capture a summary of the low-energy orientations of structural subunits in relation to one another.

The ECLAIR classifier is built using all five of the aforementioned feature categories: (1) biophysical properties, (2) conservation, (3) co-evolution, (4) SASA, and (5) docking (**Figure 6.2b**). From each of these categories we synthesized many variations of features using feature engineering techniques such as scaling, aggregation, and combination. Scaling refers to the use of either raw calculated values (i.e. the absolute SASA in $Å^2$) or normalized values (i.e. the Jensen-Shannon divergence of each residue normalized to the average of all positions per protein). Aggregation strategies dictate how features derived from multiple source of evidence (i.e. multiple protein structures for computing SASA) are combined to form a single feature (**Figure 6.2c**). Finally, multiple types of aggregation and/or scaling strategies can be used to produce several feature

variations of the same raw feature to train a classifier if they are shown to have an additive effect on predictive performance.

6.3.2 An ensemble classifier to reduce training bias

Because not all features are available for all residues, a single classifier trained on all of the aforementioned features would have limited applicability to full interactomes. For instance, current methods trained using structural features will be unable to predict for residues without structures, even when other features are available (de Vries and Bonvin, 2011; Kufareva et al., 2007; Liang et al., 2006; Porollo and Meller, 2007). Imputation methods can be used to fill in missing features, allowing a single machine learning algorithm to be applied to all data. For instance, imputation methods have been used to resolve gaps in micro-array data (Liew et al., 2011). However, these methods rely on two assumptions: (1) relatively high feature coverage, and (2) missing features are distributed randomly, without respect to the label of the data in the training set (i.e. whether the residue is at the interface) (Rubin, 1976; Schafer and Graham, 2002). The first assumption is clearly violated for predicting interface residues, as >50% of some structural features are missing in interactomes (**Figure 6.2b**).

Perhaps more importantly, features are not missing at random. Due to the technical challenges of crystallizing proteins, some proteins and protein regions are inherently more likely to be available in the PDB (Peng et al., 2004). Those protein regions that are available in co-crystal structures, which are used to assess true interface residues for training the algorithm, are also more likely to be available in other crystal structures used to gather structure-based features for predicting. Therefore, there will be very few examples of true interface residues in regions without structural features. A classifier trained using imputation to fill-in missing SASA features would then

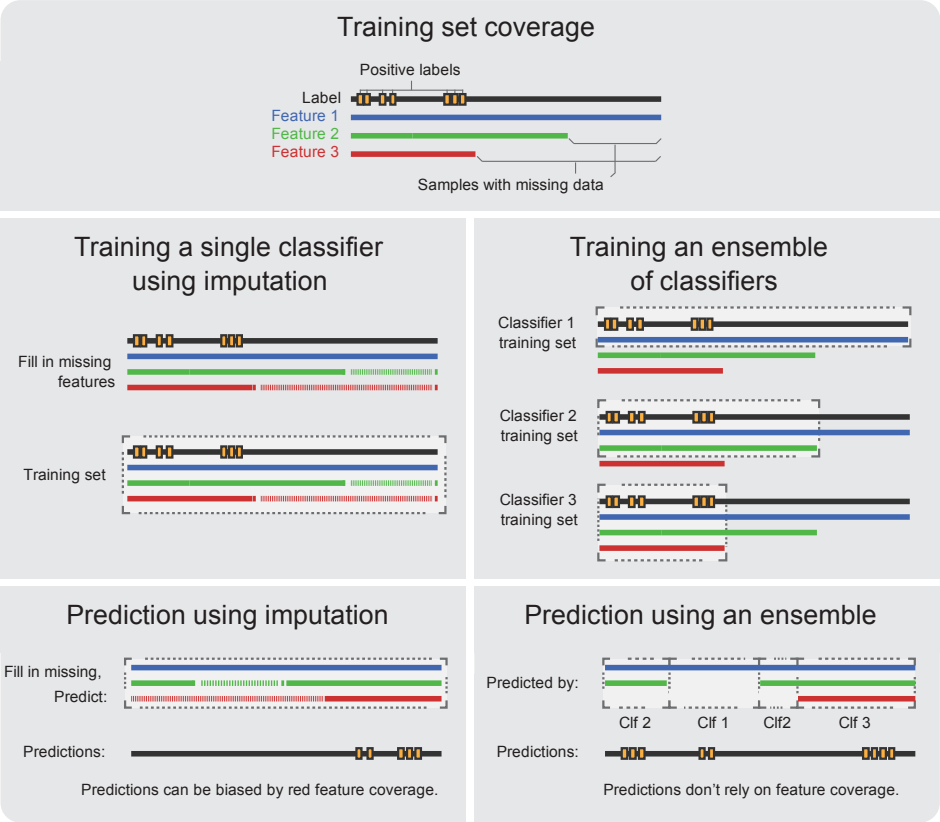

Figure 6.3 A comparison of two methods for handling missing data in classification: (1) imputation and (2) an ensemble of fully-trained classifiers. During training, imputation must fill in gaps in feature coverage, whereas an ensemble trains independent classifiers on each feature-availability scenario. Since structural feature coverage is highly correlated with the existence of known interface residues in training, imputation will fail to predict interface residues outside of regions with structural feature coverage (red). An ensemble will predict interface residues based only on the features available and will not be biased by the missing structural feature. (Figure by J. F. Beltrán)

be able to use the fact that the SASA feature is missing to always predict that the residue is not at the interface. During training, this will increase the apparent performance of the algorithm due to the classifier having learned this pattern in the training data where single crystal structures are likely to mirror the coverage of co-crystal structures; but during prediction for interactions with unknown interfaces, the algorithm will be unlikely to predict any interface residue outside of a region or full protein that has not been crystalized, which will lead to increased misclassification (**Figure 6.3**).

To address this issue, we trained an ensemble of 8 independent classifiers built on 8 subsets of features covering likely cases of feature availability. We then only use a single classifier to predict

for each residue—the classifier that was trained (without imputation) using other residues with the same set of features (**Figure 6.3**). When predicting for an interaction missing structure-based features, a classifier trained only using sequence-based features will be used, thereby allowing predictions of interface residues based only on the available features and not on the lack of structural features. In this way, the bias introduced by imputation is eliminated since missing values simply require that a different classifier in the ensemble is used.

6.3.3 Training the classifier

To train the ECLAIR ensemble classifier (**Figure 6.4a**), we first selected a training and testing set of interactions with known interface residues. We selected 400 interactions for each set, each composed evenly of homodimers and heterodimers. Importantly, we allowed no repeated proteins and no homologous interactions between the training and testing sets. We disallow repeated proteins to ensure that we don't report inflated performance metrics based upon learning shared interfaces of proteins with multiple partners. Also, since we use homology based interaction models in lieu of our classifier whenever available, our exclusion of homologous interactions ensures that our classifier performance is measured assuming no homology templates are available, thereby accurately representing the type of information that will be available for the set of interactions for which we will predict.

In the first step of training we selected a set of information-rich features from each of the five feature categories that will be used to build the final classifiers. We first applied the feature engineering techniques previously described, aggregation and scaling, on each of the raw features. We then tested combinations of engineered features in order to select a high-performing set of engineered features to represent each category. Due to their established track record of success for

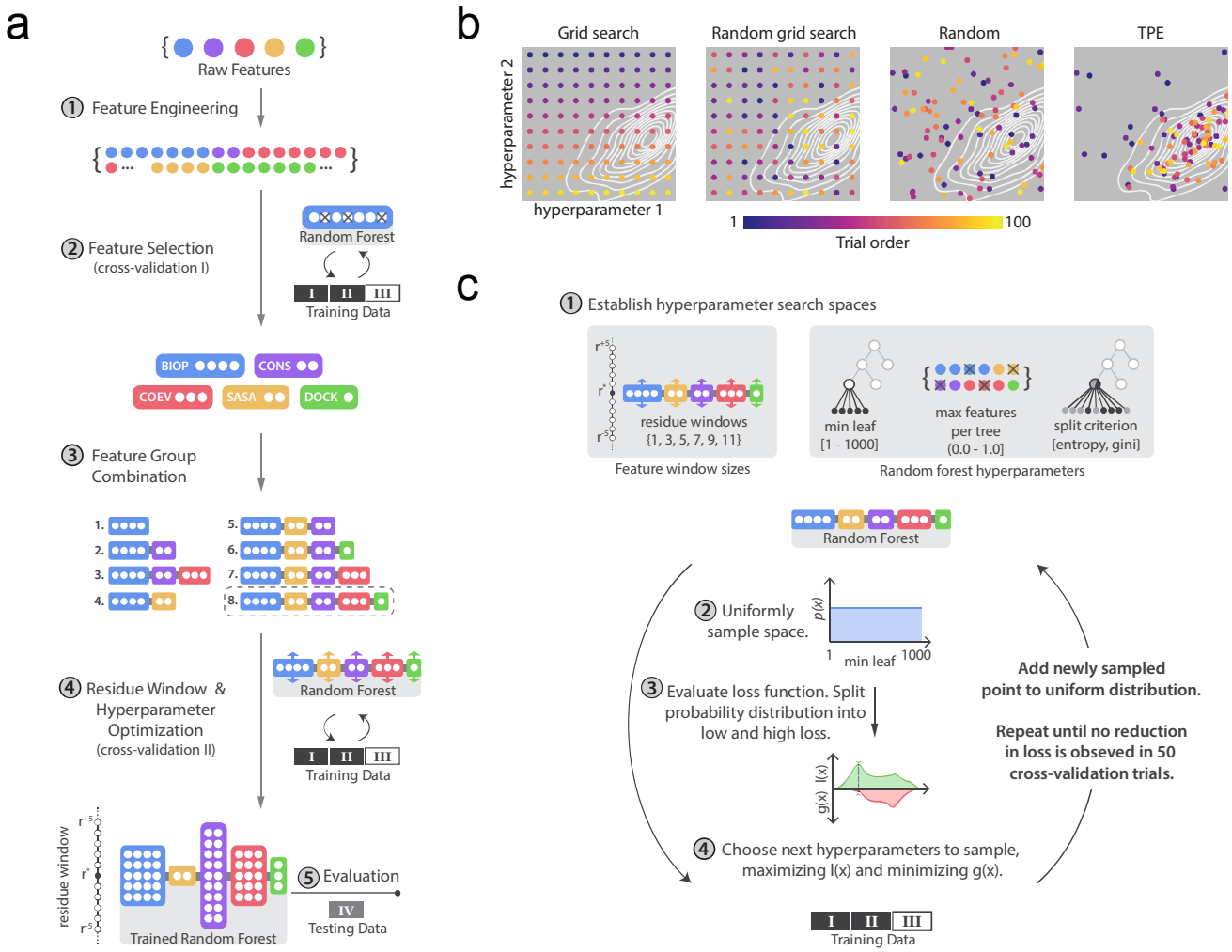

Figure 6.4 Training and optimizing the ECLAIR classifier. (a) Training the ECLAIR classifier. (b) Four methods for optimizing machine learning algorithm hyperparameters, showing the order of trials and granularity of hyperparameter sampling spaces for optimizing two hyperparameters. (c) Cross-validation strategy using TPE to optimize hyperparameters and window sizes for both feature selection and ensemble classifier training.

predicting interface residues (Esmaielbeiki et al., 2016) and in solving other biological problems (Boulesteix et al., 2012), we measured performance of random forest classifiers (Breiman, 2001) trained on each set of features and optimized over residue window sizes and algorithm hyperparameters.

As with any machine learning classifier, hyperparameters dictate the behavior of the random forest classifier and can ultimately determine its success or failure for a given prediction task. For instance, using the default parameters encoded in popular implementations of random forest vastly

undercuts performance for some classification problems (Bergstra et al., 2013), and selecting hyperparameters by hand is tedious and may introduce biases such as overfitting and cross-contamination of training and testing sets that invalidate performance metrics.

One popular automated approach for choosing hyperparameters is grid search, which, when employed properly during cross-validation, can largely avoid these pitfalls. However, grid search can be prohibitively expensive to perform at appropriate levels of granularity required to search a full hyperparameter space. Randomly ordered grid search or random continuous search may slightly improve the time needed to reach a near optimal set of hyperparameters (Bergstra and Bengio, 2012), however for hyperparameter spaces consisting of more than 2 hyperparameters being tuned concurrently, a more directed approach is needed (**Figure 6.4b**).

Recently there have been advances in Bayesian methods for selecting optimal hyperparameters at lower computational cost (Bergstra et al., 2011; Snoek et al., 2012). During cross-validation, these methods begin by sampling a hyperparameter space according to a pre-defined random distribution (i.e. normal or uniform) and then selectively sample areas of the hyperparameter space that minimize a cost function. We used a recent method, the tree-structured Parzen estimator approach (TPE) (Bergstra et al., 2011), which allowed us to simultaneously tune up to 8 hyperparameters for each random forest, including the size of residue windows over which features are included (**Figure 6.4c**).

We selected combinations of engineered features to include in final classifiers by training a set of preliminary classifiers on candidate combinations of features. We used TPE to optimize each of the preliminary classifiers, and selected the features from the highest performing classifier to represent each of the 5 feature categories (**Figure 6.5a**). In this procedure (cross validation I),

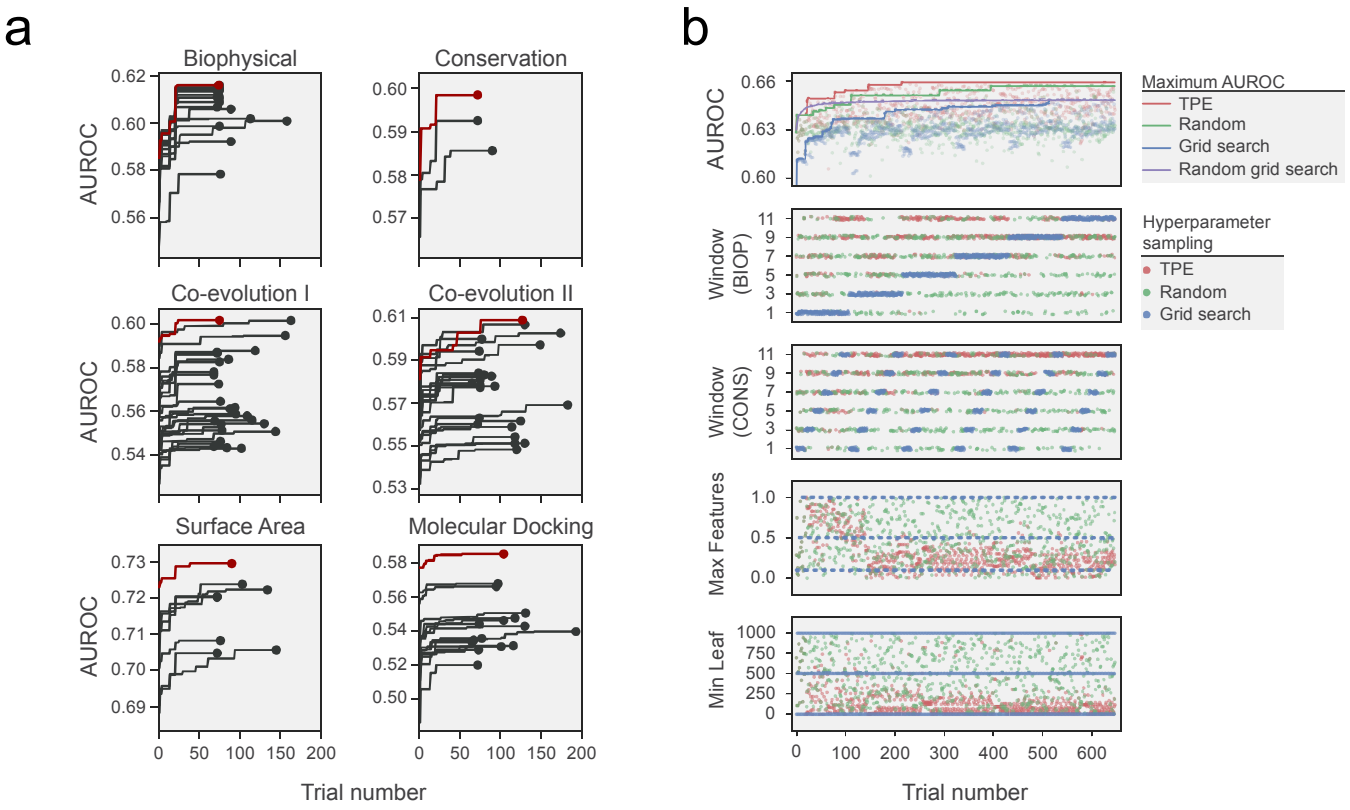

Figure 6.5 (a) Cross-validation results using TPE trials to select top performing feature or set of features (in red) in each feature category. (b) Comparison of four hyperparameter optimization methods' performance (top panel) and hyperparameter and residue window sampling patterns (bottom panels) on one of the eight sub-classifiers of the ECLAIR ensemble.

three-fold cross validation on the training set of interactions was used to evaluate each candidate set of hyperparameters, with a loss function calculated based on the area under the receiver operating characteristic (ROC) curve for the left out folds.

We then assembled the final ensemble of classifiers, which is designed to cover common feature availability scenarios and therefore allow prediction on any interaction using all available features without introducing bias (as previously described). In total, the ensemble contains 8 random forest classifiers built from the top-scoring features from cross validation I. During training of each of these 8 classifiers, cross-validation was again performed with TPE, allowing hyperparameters and window sizes to be set to their optimal values, separate of those found in the previous round of cross-validation (**Figure 6.4a**).

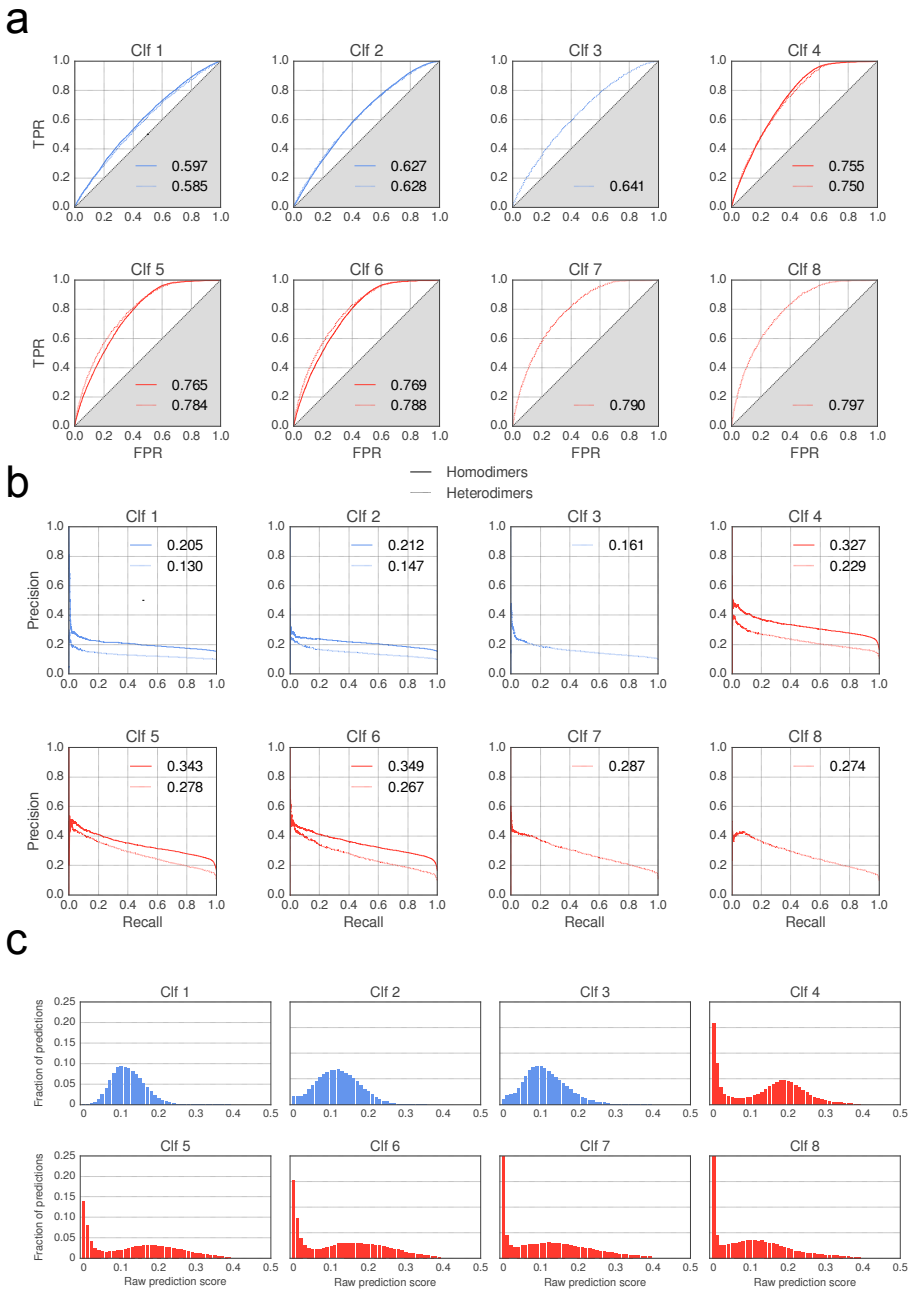

Figure 6.6 Performance of ECLAIR sub-classifiers on testing set. (a) Receiver operating characteristic (ROC) curves for each sub-classifier. (b) Precision-recall curves for each sub-classifier. (c) Distribution of raw prediction scores for each sub-classifier. For all panels, sub-classifiers plotted in blue use only sequence-based features; sub-classifiers in red used additional structure-based features.

For a single classifier of the ensemble (Classifier 2, trained on biophysical and conservation-based features), we also performed the three other aforementioned strategies for hyperparameter tuning (**Figure 6.5b**). Importantly, for sampling a minimal set of 3 potential values for each of the random forest hyperparameters, this required testing 648 combinations of hyperparameters for

grid search-based strategies. To optimize Classifier 8 (trained on all 5 feature categories) with a more reasonable (though still restrictive) 10 potential values for each random forest hyperparameter would require >1.5 million rounds of cross-validation. Not only does TPE find a better combination of hyperparameters in fewer trials, it is able to sample the full continuous range of each hyperparameter, making it the only feasible solution for searching higher-dimensional search spaces across many classifiers.

6.3.4 Evaluation of the ensemble

The performance of each of the individual 8 sub-classifiers was evaluated on the previously untouched 400 interactions in the testing set. As expected, the performances of the sub-classifiers increase as more informative features are added, with area under the ROC curve ranging from ~0.6 for the classifier trained on the fewest features to ~0.8 for the classifier trained on all features (**Figure 6.6**).

We next devised a combined score, wherein each residue was predicted by only the top classifier to which it was amenable and raw prediction values were placed on a 0-1 scale. This more closely represents the use case of the ensemble classifier as each residue in our prediction set is predicted only by the most preferred sub-classifier whose feature requirements are met. The full range of prediction scores was then divided evenly into 5 categories of interface residue propensity, with predictions in each range evaluated for their positive predictive value (precision). We find that each higher tier of predictions outperforms the previous for predicting partner-specific interface residues (**Table 6.1**). Furthermore, all tiers perform better when predicting protein-specific interface residues, highlighting the difficulty of predicting partner-specific interfaces and that

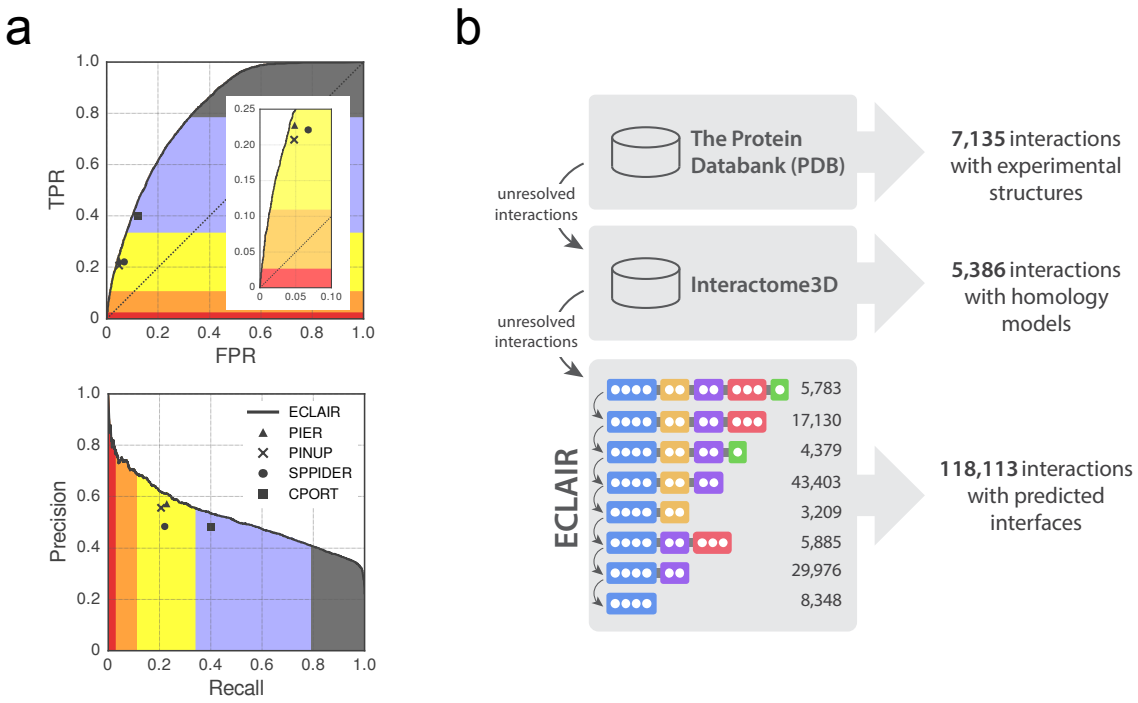

Figure 6.7 (a) ROC and precision-recall curves comparing ECLAIR with other popular interface residue prediction methods. (b) Workflow for classifying interfaces for all interactions in 8 species.

ECLAIR, even though it has been trained to predict partner-specific interfaces, still excels at predicting interfaces on a per-protein basis.

We benchmarked ECLAIR against several popular interface residue prediction methods (de Vries and Bonvin, 2011; Kufareva et al., 2007; Liang et al., 2006; Porollo and Meller, 2007) by executing available prediction servers on our testing set. We find that ECLAIR outperforms these servers in precision, recall, and false positive rate (**Figure 6.7a**). Furthermore, ECLAIR provides a continuous range of predictions, whereby the user can select their own tolerance for tradeoffs between these metrics. For instance, in designing targeted mutagenesis experiments, a researcher may choose a small number of targets using prediction scores from the High or Very High interface potential categories.

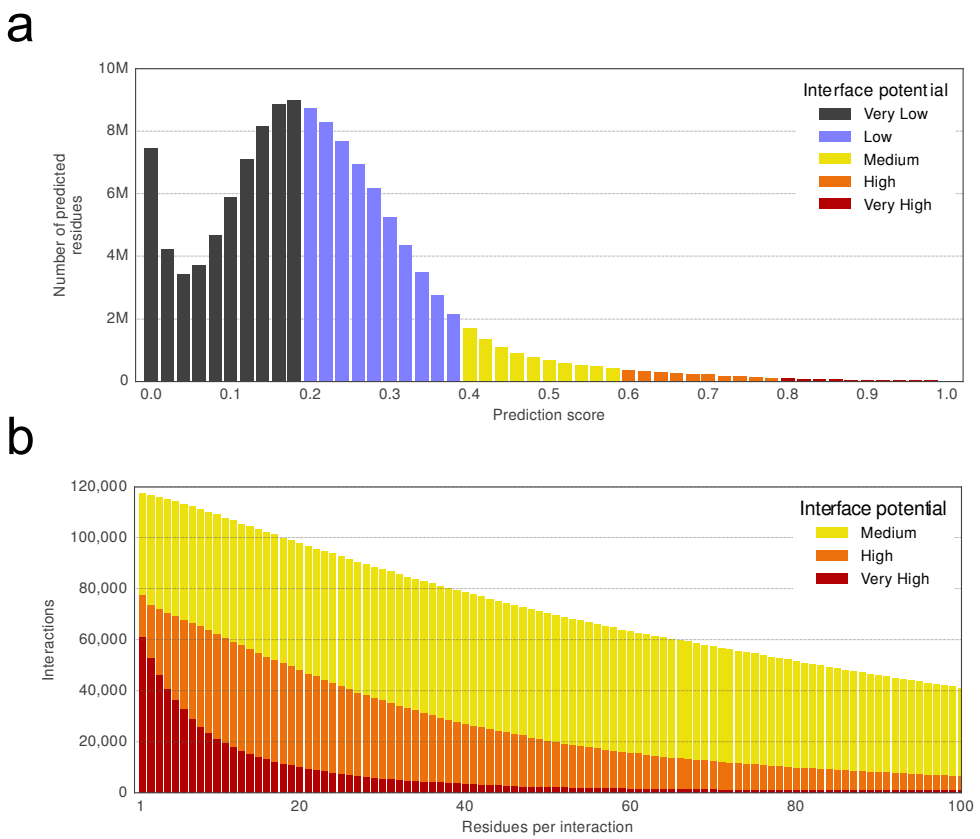

Figure 6.8 (a) The distribution of prediction scores across five confidence categories for all ECLAIR predictions in 8 organisms. (b) Cumulative distribution of interactions with ≥ n residues classified as interface for each of the highest interface potential categories.

6.3.5 Classification of unknown interfaces with systematic computational and experimental evaluation

After performance evaluation, the 8 sub-classifiers were finally retrained on the entire set of known interface residues, including both the training and testing sets. The fully trained ensemble classifier was used to predict interfaces residues in each of 118,113 interactions without known experimental structures or homology models (**Fig 6.7b**). As expected, the majority of residues are predicted to be away from the interface (only ~12% of residues in our training set are at the interface), and appropriately few residues are predicted with High and Very High confidence (**Figure 6.8a**). Importantly, the distribution of predicted interface residues ensures good coverage of the interactomes; >50% of interactions have at least 10 residues predicted with High or Very High

confidence to be at the interface, and >99% interactions have at least one Medium confidence prediction, making the dataset widely applicable to all interactions (**Figure 6.8b**).

We next investigated the functional properties of predicted interface residues to judge their use in genomic-scale studies. For instance, if it can be shown that predicted interface residues share properties of known interface residues, then it may be possible to use predicted interface residues in genomic-scale analyses.

First, we analyzed the properties of interface residues in relation to human disease. It has been shown that disease mutations are enriched at the interface of interacting proteins (Wang et al., 2012), suggesting that disruption of binding with one or more partners may contribute to disease. However, >40% of known missense and nonsense human disease mutations cause alterations to proteins without structurally resolved interaction interfaces for any of their known binding partners. We find that disease mutations also preferentially occur at predicted interface residues, at similar rates to known interface residues in PDB co-crystal structures (**Figure 6.9a**), indicating the viability of using predicted interfaces to study molecular disease mechanisms. Furthermore, each more confident bin of predicted interface residues is more likely to contain disease mutations than the previous, showing that ECLAIR prediction scores are correlated with true protein function.

We also performed a massive mutagenesis experiment to test whether mutations at interaction interfaces are more likely to disrupt interactions than mutations away from the interface. Using a yeast two-hybrid reporter, we were also able to assess effects of mutations on multiple binding partners. Even for this relatively more complex case, we find that the disruption rates for mutations at known interface residues is quite similar to disruption rates for mutations of predicted interface

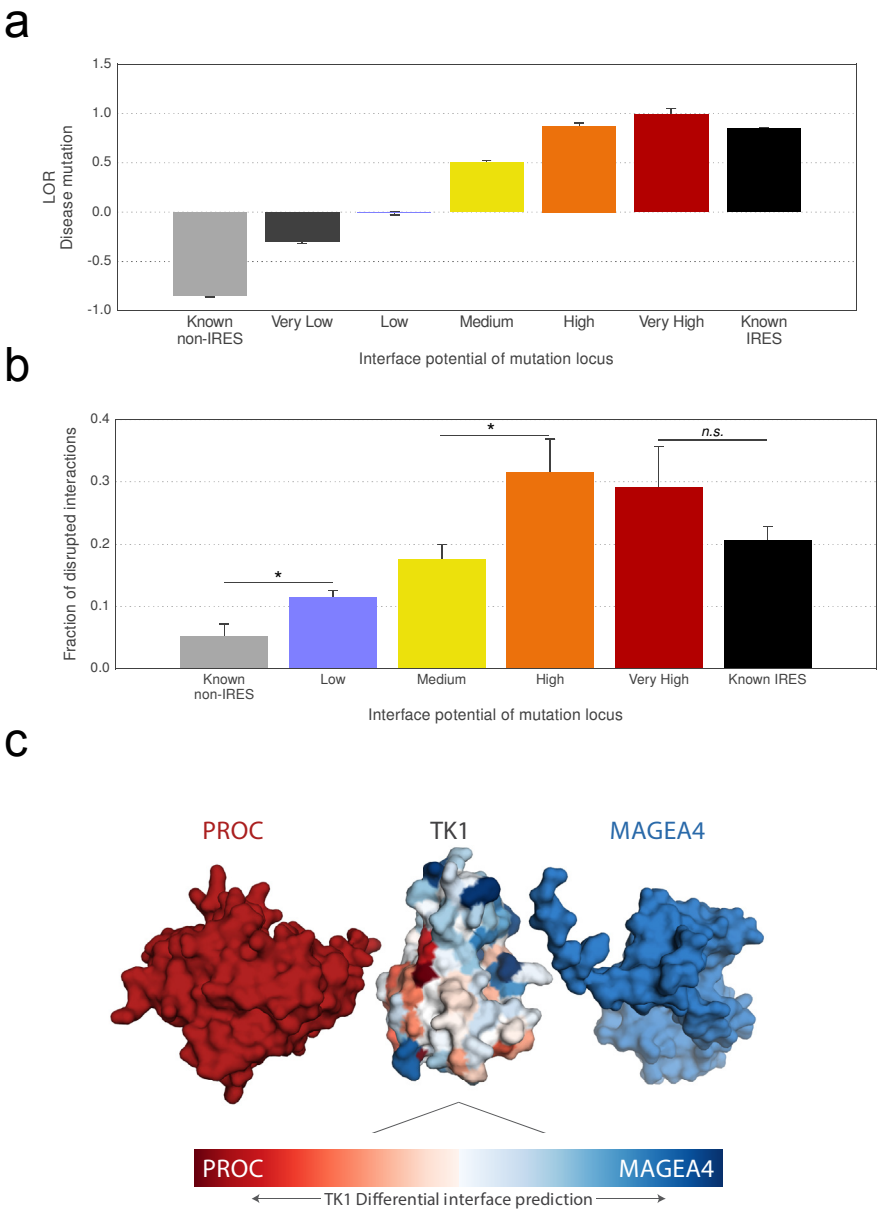

Figure 6.9 Functional interface prediction analyses. (a) Log odds ratio of disease mutation being at the interface vs. any residue being at the interface, for both known and predicted interfaces. (b) Fraction of interactions disrupted by the introduction of random population variants in known and predicted interfaces (* denotes significant ($p < 0.05$); n.s. denotes not significant by a Z-test). (c) Superimposed docking results of two different partners with TK1. The differentially predicted interfaces of TK1 with each of its partners corresponds with the orientation of the docked poses. For (a) and (b), measurements are shown ± SE.

residues (**Figure 6.9b**). Furthermore, even mutations of residues with a 'Low' predicted interface potential are significantly more likely to disrupt interactions than mutations of residues known to be away from the interface. This suggests that there is viable functional signal in ECLAIR predictions even in the 'Low' potential category.

To show how partner-specific features can affect interface predictions in a binding partner-dependent manner, we highlight an interaction whose predicted interfaces are strongly influenced by a single partner-specific feature. In the example, the protein TK1 is shown with the docking result with each of two partners, PROC and MAGEA4 (**Figure 6.9c**). We note that the predicted interface residues on TK1 are drastically different for each partner, and that the areas with elevated interface potential correspond to the position of the two docking results, showing how differential partner-specific features can lead to differential interface predictions.

6.4 DISCUSSION

ECLAIR is a genomics-era biological resource that leverages multiple sources of structural information and a new ensemble machine learning algorithm to provide the most complete and reliable map of protein interaction interfaces to date. While previous methods have sacrificed either quality or scale, ECLAIR makes use of the most predictive available features for any given interaction, allowing it to make interactome-wide predictions amenable to genome-scale analyses. The sheer scale of ECLAIR (>100 million residue predictions for 118,113 interactions in 8 organisms) and ease of accessibility (http://eclair.yulab.org) make it a truly unique resource likely to be applied in genomic-scale surveys and targeted studies.

With future increases to the scale of biological databases from which we derive features, we expect that ECLAIR will come to encompass even higher confidence predictions for many more interactions. This may address some limitations of structural databases today, for instance the PDB is depleted of disordered proteins (Peng et al., 2004), and it has been shown that disordered regions can form interfaces (Dunker et al., 2008). Since our classifier has not been trained on disordered interfaces, it is unlikely to predict new disordered interfaces. However, the ensemble classifier

structure of ECLAIR uniquely positions it to incorporate all newly-available evidence into interface predictions without sacrificing quality or scale, ensuring the highest quality map of interaction interfaces now and in the future.

We have performed analyses showing the intersection of this new dataset with known disease mutations, and anticipate it will help to bridge the divide between genomic-scale datasets and structural proteomic analyses. Now that sequencing data from many contexts is readily available, for instance from population variant studies (Fu et al., 2013; Lek et al., 2016) and cancer studies (Forbes et al., 2011; Kandoth et al., 2013), researchers have become increasingly interested in ways to assess the potential functional consequences of variants on a genomic scale (Cingolani et al., 2012; Hofree et al., 2016; Lawrence et al., 2013; Tasan et al., 2015). For instance, recently we and others have developed methods to predict functional cancer driver mutations by finding hotspots of mutations in the structural proteome (Kamburov et al., 2015; Meyer et al., 2016; Yang et al., 2015). With the pan-interactome map of protein interfaces presented, we can now go a step further to predict specific etiologies of cancer and disease based on induced biophysical effects (Kucukkal et al., 2015; Li et al., 2014) that may break interactions. Because our interface map is partner-specific, it can also be applied to predict pleiotropic effects, wherein several mutations in a single protein may affect different pathways depending upon which binding interfaces are mutated (Wang et al., 2012). Importantly, this could be used in rational drug design to selectively target specific protein functional sites (Lounnas et al., 2013).

Finally, the tiered ensemble form of ECLAIR represents a broadly applicable paradigm in practical machine learning for the observational sciences. Specifically, the approach we present is well-suited to solving problems with large amounts of non-uniformly missing data, which very frequently occur in biology due to study biases. Furthermore, we have shown that hyperparameter

optimization, which we used to independently tune each classifier of the ensemble, can drastically improve the performance of classifiers. For both existing and future biological classification methods, it may be possible to improve performance through more consistent application of hyperparameter optimization methods, which is surprisingly lacking in much of the current literature.

Interface Potential	Partner-specific PPV	Partner-specific FPR	Per-protein PPV	Per-protein FPR
Very High	0.587	0.002	0.818	0.001
High	0.569	0.015	0.752	0.011
Medium	0.474	0.106	0.645	0.090
Low	0.362	0.408	0.518	0.378
Very Low	0.278	1.000	0.411	1.000

Table 6.1. Positive predictive value (PPV, within ± 3 residues) and false positive rate (FPR) of predicted interface residues. Categories are inclusive of higher-confidence categories.

6.5 REFERENCES

Arabidopsis Interactome Mapping, C. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Bergstra, J., and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research *13*, 281-305.

Bergstra, J., Yamins, D., and Cox, D.D. (2013). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. Paper presented at: Proceedings of the 12th Python in Science Conference (Citeseer).

Bergstra, J.S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. Paper presented at: Advances in Neural Information Processing Systems.

Boulesteix, A.L., Janitza, S., Kruppa, J., and Konig, I.R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wires Data Min Knowl *2*, 493-507.

Breiman, L. (2001). Random Forests. Mach Learn *45*, 5-32.

Brunk, E., Mih, N., Monk, J., Zhang, Z., O'Brien, E.J., Bliven, S.E., Chen, K., Chang, R.L., Bourne, P.E., and Palsson, B.O. (2016). Systems biology of the structural proteome. BMC Syst Biol *10*, 26.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92.

Das, J., Fragoza, R., Lee, H.R., Cordero, N.A., Guo, Y., Meyer, M.J., Vo, T.V., Wang, X., and Yu, H. (2014). Exploring mechanisms of human disease through structurally resolved protein interactome networks. Mol Biosyst *10*, 9-17.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

de Vries, S.J., and Bonvin, A.M. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PloS one *6*, e17695.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., and Uversky, V.N. (2008). The unfoldomics decade: an update on intrinsically disordered proteins. BMC genomics *9 Suppl 2*, S1.

Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.C., and Deane, C.M. (2016). Progress and challenges in predicting protein interfaces. Brief Bioinform *17*, 117-131.

Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., and Tress, M.L. (2009). Progress and challenges in predicting protein-protein interaction sites. Brief Bioinform *10*, 233-246.

Forbes, S., Bindal, N., Bamford, S., Cole, C., Kok, C., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research *39*, 50.

Fu, W., O'Connor, T., Jun, G., Kang, H., Abecasis, G., Leal, S., Gabriel, S., Rieder, M., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Guharoy, M., and Chakrabarti, P. (2010). Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics *11*, 286.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins *47*, 409-443.

Hofree, M., Carter, H., Kreisberg, J.F., Bandyopadhyay, S., Mischel, P.S., Friend, S., and Ideker, T. (2016). Challenges in identifying cancer genes by analysis of exome sequencing data. Nature communications *7*, 12096.

Hopf, T.A., Scharfe, C.P., Rodrigues, J.P., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife *3*.

Hwang, H., Vreven, T., and Weng, Z. (2014). Binding interface prediction by combining protein-protein docking results. Proteins *82*, 57-66.

Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proceedings of the National Academy of Sciences of the United States of America *112*, E5486-5495.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Kim, P.M., Lu, L.J., Xia, Y., and Gerstein, M.B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. Science *314*, 1938-1941.

Kucukkal, T.G., Petukh, M., Li, L., and Alexov, E. (2015). Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Current opinion in structural biology *32*, 18-24.

Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007). PIER: protein interface recognition for structural proteomics. Proteins *67*, 400-417.

Kuhlbrandt, W. (2014). Cryo-EM enters a new era. eLife *3*, e03678.

Lawrence, M., Stojanov, P., Polak, P., Kryukov, G., Cibulskis, K., Sivachenko, A., Carter, S., Stewart, C., Mermel, C., Roberts, S.*, et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Lensink, M.F., Velankar, S., Kryshtafovych, A., Huang, S.Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R.*, et al.* (2016). Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins.

Li, M., Petukh, M., Alexov, E., and Panchenko, A.R. (2014). Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. J Chem Theory Comput *10*, 1770-1780.

Liang, S., Zhang, C., Liu, S., and Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. Nucleic Acids Res *34*, 3698-3707.

Liew, A.W., Law, N.F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief Bioinform *12*, 498-513.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295-299.

Lounnas, V., Ritschel, T., Kelder, J., McGuire, R., Bywater, R.P., and Foloppe, N. (2013). Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. Computational and structural biotechnology journal *5*, e201302011.

Meyer, M.J., Lapcevic, R., Romero, A.E., Yoon, M., Das, J., Beltran, J.F., Mort, M., Stenson, P.D., Cooper, D.N., Paccanaro, A.*, et al.* (2016). mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum Mutat *37*, 447-456.

Morcos, F., Hwa, T., Onuchic, J.N., and Weigt, M. (2014). Direct coupling analysis for protein contact prediction. Methods Mol Biol *1137*, 55-70.

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nature methods *10*, 47-53.

Peng, K., Obradovic, Z., and Vucetic, S. (2004). Exploring bias in the Protein Data Bank using contrast classifiers. Pac Symp Biocomput, 435-446.

Pieper, U., Webb, B., Barkan, D., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E., Pettersen, E., Huang, C.*, et al.* (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research *39*, 74.

Porollo, A., and Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. Proteins *66*, 630-645.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212-1226.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika *63*, 581-590.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y.*, et al.* (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647-660.

Schafer, J.L., and Graham, J.W. (2002). Missing data: our view of the state of the art. Psychological methods *7*, 147-177.

Snoek, J., Larochelle, H., and Adams, R.P. (2012). Practical bayesian optimization of machine learning algorithms. Paper presented at: Advances in neural information processing systems.

Tasan, M., Musso, G., Hao, T., Vidal, M., MacRae, C.A., and Roth, F.P. (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nat Methods *12*, 154-159.

Vakser, I.A. (2013). Low-resolution structural modeling of protein interactome. Current opinion in structural biology *23*, 198-205.

Vo, T.V., Das, J., Meyer, M.J., Cordero, N.A., Akturk, N., Wei, X., Fair, B.J., Degatano, A.G., Fragoza, R., Liu, L.G.*, et al.* (2016). A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. Cell *164*, 310-323.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.

Xie, L., Ge, X., Tan, H., Xie, L., Zhang, Y., Hart, T., Yang, X., and Bourne, P.E. (2014). Towards structural systems pharmacology to study complex diseases and personalized medicine. PLoS computational biology *10*, e1003554.

Yang, F., Petsalaki, E., Rolland, T., Hill, D.E., Vidal, M., and Roth, F.P. (2015). Protein domain-level landscape of cancer-type-specific somatic mutations. PLoS computational biology *11*, e1004147.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Zhou, H.X., and Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics *23*, 2203-2209.

APPENDIX A

Supplementary Information for:

A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human

A.1 Detection rates of the PRS and NRS

Of the 168 NRS pairs, 78 are between proteins with different sub-cellular localizations and 90 have the same sub-cellular localization (Matsuyama et al., 2006). However, there is no significant difference in the fraction of random pairs detected by either Y2H (0/78 and 0/90 for the two sets respectively, P=0.92 using a Z test) or PCA (8/78 and 9/90 for the 2 sets respectively, P=0.96 using a Z test).

To examine if there are any species-specific biases of our Y2H assay, we computed the fractions of PRS and PRS-nonY2H (subset of PRS interactions that have been detected using an assay other than Y2H) interactions in the three different species that are recapitulated by our Y2H assay. We find that there is no significant difference between the detection rates across species (Figure A.1a; P>0.35 for all pairs, Z test). Furthermore, we find that there is no significant difference in interaction density (*i.e.*, number of interactions detected divided by total number of protein pairs screened) for FissionNet and previously reported Y2H interactomes in *S. cerevisiae* (Yu et al., 2008) and human (Rolland et al., 2014) (Figure A.1b; all interaction densities differ by <2 fold). These results confirm that our Y2H assay has no species-specific detection biases.

A.2 Calculating the coexpression of genes

To measure the coexpression of transcripts corresponding to proteins involved in FissionNet interactions, we calculated the Pearson Correlation Coefficient (*PCC*) between their expression profiles: expression values measured at different time-points in the cell cycle (Rustici et al., 2004). We also calculated the *PCC* between expression profiles of transcripts corresponding to proteins involved in high-quality *S. pombe* interactions from literature curation. Finally, we defined two different sets of random pairs: (1) all random pairs, (2) random pairs by permuting edges between proteins in the network. We first compared the different distributions using a KS test. Next, we calculated the fractions of significantly co-expressed interactions, as well as the fraction of significantly co-expressed random pairs. We defined significant coexpression as *PCC* ≥a threshold value. When comparing the fractions of interactions or pairs that are significantly co-expressed, *P*-values were calculated using a *Z* test.

A.3 Other functional properties of FissionNet

For other calculations, since small-scale studies could focus on proteins with more complete annotations in GO (Das and Yu, 2012), we restricted our analyses to a set of proteins found in both high-quality literature-curated *S. pombe* interactions (Das and Yu, 2012) and interactions in FissionNet. We then defined 3 sets of protein pairs such that both proteins are from the previously defined set: (1) high-quality *S. pombe* interactions from literature curation, (2) *S. pombe* interactions from FissionNet, and (3) all pairs of proteins for which the two proteins have never been reported to interact. We performed the following calculations on these 3 sets:

A.3.1 Calculating functional similarity

We calculated functional similarity using a total ancestry method that computes all pairwise functional similarities in a set of proteins by determining for each given pair of proteins, the number of other protein pairs sharing the same set of parent GO terms (Yu et al., 2007). In this framework, a pair of proteins that are very dissimilar will share their GO ancestry with a large number of other protein pairs. Conversely, a pair of proteins that are very similar will share their GO ancestry with only a few or none of the other pairwise combinations of proteins in the same set. Each similarity score for a pair of proteins was computed as a percentile ranking of their total ancestry score among all such scores calculated for all pairwise combinations of proteins in the set. We considered the top 1% of protein pairs in this ranking to be functionally similar. *P*-values were calculated using a *Z* test.

A.3.2 Calculating co-localization

To calculate the co-localization of proteins involved in FissionNet interactions, we calculated the fraction of protein pairs that have the same sub-cellular localization (Matsuyama et al., 2006). *P*-values were calculated using a *Z* test.

A.4 Conservation of genes

To analyze the extent to which genes are conserved, we calculated the fraction of genes in the reference species *i* that also have orthologs in the other species *j*:

$$Gene\_cons_{ij} = \frac{G_i^{j'}}{G_i}$$

where G_i denotes the total number of genes in species i and $G_i^{j'}$ the number of genes in species *i* that have corresponding orthologs in species *j*. Using ortholog annotations from PomBase and the

Saccharomyces Genome Database, we computed the extent of gene conservation between different species pairs for all coding genes (Figure A.2a) (Cherry et al., 2012; McDowall et al., 2015). We also used orthologs from InParanoid to compute the extent of gene conservation between different species pairs for all coding genes (Figure A.2b) (Sonnhammer and Ostlund, 2015). We observe the same gene conservation trends regardless of which database is used for determining orthology, confirming the robustness of our result.

A.5 Estimating true interaction conservation fractions

To calculate the extent to which interactions are conserved, we focused only on those interactions that can be conserved, *i.e.*, both proteins involved in the interaction have orthologs in the other species. For each pair of organisms, we used both organisms as the reference (six comparisons for 3 species). We mapped interactions in the reference species to their corresponding ortholog pairs in the other species and tested these pairs using our Y2H assay in a pairwise fashion. We performed pairwise retests because we have shown earlier that not all interactions detected by Y2H in a pairwise fashion will be detected in a high-throughput screen where individual baits are tested against minipools of ~188 preys (Yu et al., 2008). Overall, results from these pairwise retests for all three species (a total of ~20,000 individual Y2H experiments) are used to obtain the observed conservation fraction. To accurately estimate the true conservation fraction, we used a rigorous Bayesian framework that takes into account both the false positive and false negative rates of our Y2H assay, and computes the true conservation fraction from the observed fraction.

Using the law of total probability, we can write:

$$P(D \mid I') = P(D \mid I,\ I') \times P(I \mid I') + P(D \mid \bar{I},\ I') \times P(\bar{I} \mid I') \quad (1)$$

Here, I' denotes the event that an interaction occurs in the reference species, I the event that the interaction occurs in another species, $\bar{I}$ the event that the interaction does not occur in the other species and D the event that it is detected in the other species using our Y2H pipeline. The observed conservation rate is $P(D|I')$. The true conservation rate is $P(I|I')$. As an interaction in the reference species can only be either conserved or rewired in the other species:

$$P(I \mid I') + P(\bar{I} \mid I') = 1 \quad (2)$$

Finally, we can assume conditional independence between D and I' given I. In other words, given that an interaction occurs in the other species, whether it is Y2H detectable in that species and whether its ortholog pair interacts in the reference species are independent of each other. Using this:

$$P(D \mid I, I') = \frac{P(D,I,I')}{P(I,I')} = \frac{P(D, I' \mid I) \times P(I)}{P(I,I')} = P(D \mid I) \times \left(\frac{P(I' \mid I) \times P(I)}{P(I,I')} \right) = P(D \mid I) \quad (3)$$

Using similar arguments,

$$P(D \mid \bar{I}, I') = P(D \mid \bar{I}) \quad (4)$$

Substituting equations (2), (3) and (4) in equation (1), we obtain:

$$P(I \mid I') = \frac{P(D \mid I') - P(D \mid \bar{I})}{P(D \mid I) - P(D \mid \bar{I})} \quad (5)$$

$P(D|I')$ is estimated using the fraction of interactions in the reference species that are detected by Y2H to interact in the other species (f_d). $P(D|I)$ is estimated using the fraction of a set of true interactions (PRS) that we can detect using our Y2H assay (f_{prs}). Finally, $P(D|\bar{I})$ is estimated using the fraction of a set of random pairs that are unlikely to interact (NRS) that we can detect using our Y2H assay (f_{nrs}). So, for any species pairs:

$$P(I \mid I') = \frac{f_d - f_{nrs}}{f_{prs} - f_{nrs}} \quad (6)$$

We can estimate the error using the delta method:

$$SE_{P(I|I')} = \sqrt{\frac{\left(f_{prs}-f_{nrs}\right)^2 \times \left(SE_{f_d}\right)^2 + \left(f_{prs}-f_d\right)^2 \times \left(SE_{f_{nrs}}\right)^2 + \left(f_d-f_{nrs}\right)^2 \times \left(SE_{f_{prs}}\right)^2}{\left(f_{prs}-f_{nrs}\right)^4}} \quad (7)$$

A.6 Interaction conservation using assays other than Y2H

We examined the observed conservation as detected by individual assays rather than using overall interactome networks from the literature as these are derived from assays with varied and unknown false positive and false negative rates. However, for a single assay with unknown false positive and false negative rates, while we will be unable to calculate the true underlying conservation fraction, we can still compute the observed conservation fraction. We first calculated the fraction of FissionNet interactions whose corresponding *S. cerevisiae* and human ortholog pairs have been shown to interact in co-crystal structures (Das and Yu, 2012). We find that that fission yeast interactions are better conserved in human than in budding yeast (Figure A.3a; >2 fold difference in observed conservation, $P<10^{-3}$). Next, we calculated the fraction of FissionNet interactions whose corresponding *S. cerevisiae* and human ortholog pairs have been detected as interacting by proteome-scale affinity purification/mass spectrometry experiments (Gavin et al., 2006; Huttlin et al., 2015; Krogan et al., 2006). Here, we also find that fission yeast interactions are better conserved in human than in budding yeast (Figures A.3b and A.3c; >1.5 fold difference in observed conservation, $P<10^{-3}$ in both cases).

A.7 Identifying proteins conserved in eukaryotes

To identify proteins that are conserved across eukaryotes, we used clusters of conserved eukaryotic orthologous groups of genes (KOGs) as defined by Koonin *et al.* (Koonin et al., 2004). These

conserved KOGs often comprise genes essential for survival and could be considered to approximate "a minimal set of essential eukaryotic genes" (Koonin et al., 2004). Each KOG consists of orthologous genes in up to 7 representative eukaryotic species studied by the authors. We defined proteins conserved in eukaryotes as those proteins from these KOGs that are conserved in ≥5 species.

A.8 Interaction conservation in different biological processes

We used the Gene Ontology (GO) (Ashburner et al., 2000) to categorize interactions based on the annotations of the proteins involved. We computed interaction conservation in GO Slim Biological Process (BP) categories, a set of 70 terms representative of diverse biological processes not specific to any one organism. For all analyses, we considered only genes annotated with experimental evidence codes (Ashburner et al., 2000). We considered an interaction to be within a category if either of its interacting proteins is annotated in that category or one of its children.

A.9 Sequence conservation of proteins and interactions

To determine the sequence conservation between two proteins, alignments were produced using the pairwise2.align.global function of the BioPython Python module, an implementation of the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970). We used the BLOSUM62 scoring matrix, a gap-open penalty of -10 and a gap-extend penalty of -0.5. Two amino acids are considered similar if the BLOSUM62 score associated with a substitution between the two residues is >0. Unless otherwise specified, sequence similarity is measured with sequences of *S. pombe* proteins serving as the reference. Sequence similarity between an *S. pombe* protein and an ortholog in another species is measured as the fraction of *S. pombe* residues similar to their aligned residues in either *S. cerevisiae* or human. To calculate the sequence similarity of pairs of proteins

with orthologous pairs, the individual sequence similarities of each protein with their orthologs are averaged. *P*-values were calculated using a *Z* test.

A.10 Interface domain conservation based on co-crystal structures

We compiled a set of co-crystal structures from the PDB representing human protein-protein interactions. For each structure, we calculated interface residues using NACCESS to determine surface residues whose solvent accessible surface area was altered by ≥1Å^2 between bound and unbound states (Hubbard, 1996). To determine interface residues of protein interactions, we took the union of interface residues determined from each representative PDB chain pair for which at least 5 interface residues were calculated in each chain. In the human interactions, we identified Pfam domains at the interaction interface as those domains containing at least 5 interface residues. All domains not meeting this criterion are considered 'Other' as we don't know if they facilitate the interaction or not. We then aligned the full human protein sequences in each interaction to their orthologs in *S. pombe* and *S. cerevisiae* using the alignment method mentioned previously. Here, we used the human sequences as the reference and only calculated sequence similarity within the portions of the alignment in the human domain regions. *P*-values were calculated using a *U* test.

A.11 ClusterOne

We performed clustering with ClusterONE (Nepusz et al., 2012). ClusterONE finds overlapping functional modules and is specifically tuned for clustering biological networks. We used ClusterONE with parameters *s*=3 (minimum cluster size) and *d*=0.5 (minimum cluster density) and found 193 clusters in our network. Since proteins can belong to multiple clusters, we defined an intra-cluster interaction as any interaction for which there is a cluster that contains both proteins and an inter-

cluster interaction as any interaction for which both proteins belong to clusters, but there is no cluster that contains both proteins. Intra-cluster and inter-cluster conservations were calculated using the fraction of interactions within and across clusters that are detected as conserved using our Y2H assay, transformed via the Bayesian framework described above to obtain the true conservation fractions (Figure 4.6a). *P*-values were calculated using a *Z* test.

A.12 Affinity propagation clustering

We also used affinity propagation clustering (APC) to generate clusters from FissionNet (Frey and Dueck, 2007). APC relies only on the topological properties of the network to generate clusters. The algorithm requires a pairwise similarity measure as input. This was defined as follows:

$$Sim_{i,j} = (1 + Diam_{FN}) - Dist_{i,j}$$

Here $Dist_{i,j}$ is the graph distance between nodes *i* and *j*. $Diam_{FN}$ is the graph diameter of FissionNet, *i.e.*, the maximum distance between any two nodes. Graph distance was set to 1 + $Diam_{FN}$ for node pairs that are not connected. Thus, $Sim_{i,j}$ represents the normalized similarity between two nodes. It will take the highest value (equal to $Diam_{FN}$) for nodes that are directly connected and the lowest value (0) for nodes that are not connected at all.

Observed intra-cluster and inter-cluster conservations were obtained by calculating the fraction of interactions within and across clusters that are detected as conserved using our Y2H assay. *P*-values were calculated using a *Z* test.

A.13 Gene Ontology

We used Gene Ontology (GO) (Ashburner et al., 2000) and GO Slim annotations from PomBase (McDowall et al., 2015) to cluster FissionNet based on known biological processes. Intra-process

and inter-process conservations were calculated using the fraction of interactions within and across processes that are detected as conserved using our Y2H assay, and transformed via the Bayesian framework described above to obtain the true conservation fractions (Figure 4.6b; all GO). *P*-values were calculated using a *Z* test.

A.14 Distribution of intact and coevolved interactions across species

We computed the log-odds ratios for 3 scenarios: an interaction is intact in both species pairs (*S. pombe*-*S. cerevisiae* and *S. pombe*-human), an interaction is coevolved in both species pairs, an interaction is intact in one species pair but coevolved in the other:

$$LOR = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$$

where, p_1 is the observed fraction of interactions in each category and p_2 the expected fraction of interactions in each category. The expected fraction is calculated assuming independence between the events of being intact/coevolved in each species pair. Standard error was calculated using the delta method:

$$SE_{LOR} = \sqrt{\left(\frac{SE_{p1}{}^2}{p_1{}^2 \times (1-p_1)^2} + \frac{SE_{p2}{}^2}{p_2{}^2 \times (1-p_2)^2}\right)}$$

P-values were calculated using a *Z* test.

A.15 Direct Coupling Analysis (DCA) for coevolutionary studies

To measure inter-protein evolutionary residue correlations, we performed coevolutionary analyses using DCA, which disentangles direct from indirect correlations among residue positions in

evolutionarily-derived multiple sequence alignments (MSA). The most highly correlated residue pairs, as indicated by a high direct information (DI) score, can be used to predict contact residues, protein structures, and complex interfaces (Morcos et al., 2011). We compiled a list of orthologous protein sequences in *S. pombe*, *S. cerevisiae*, and 26 other yeast species by computing reciprocal best BLASTP hits between the proteome of *S. pombe* and the proteomes of these species accessed from UniProt (*82*). Cd-Hit was used to eliminate redundant sequences with >90% sequence identity from the list of orthologs for each *S. pombe* protein (Li and Godzik, 2006). MSAs among all remaining orthologs for each *S. pombe* protein were assembled using Clustal Omega (Sievers et al., 2011). For each studied *S. pombe* interaction known to be intact or coevolved in *S. cerevisiae*, we concatenated MSAs for the two proteins involved and ran DCA using default parameters to find the inter-protein residue pairs with the highest correlations (DI scores). Homodimers were excluded from this calculation as it is impossible to disentangle intra from inter-protein evolutionary pressures. *P*-values were calculated using a *U* test.

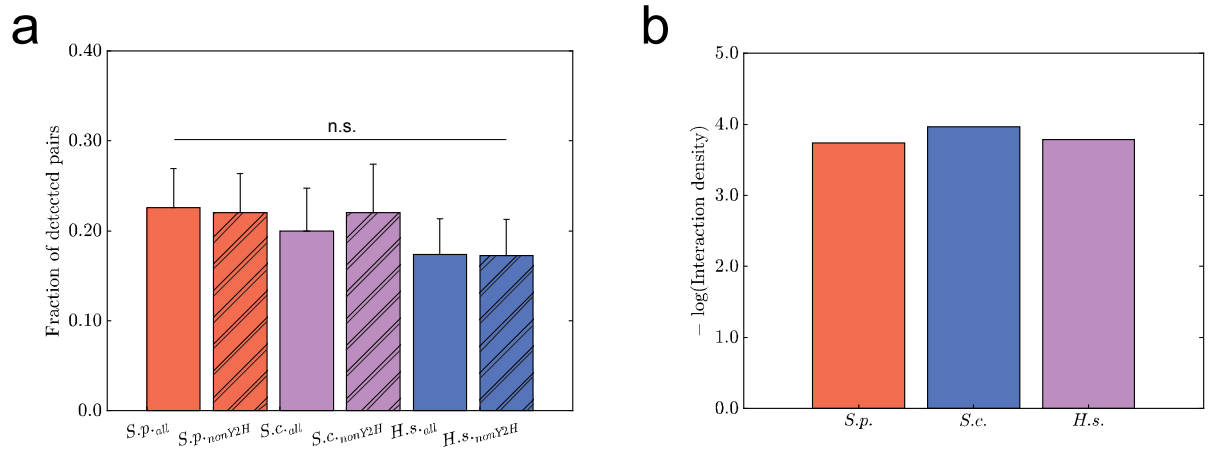

Figure A.1 (a) Y2H detection rates of PRS and PRS_nonY2H (subset of PRS interactions that have been detected using an assay other than Y2H) interactions in fission yeast, budding yeast and human. (b) Interaction density, i.e., interactions detected out of the total number of proteins pairs screened (log scale) in different organisms. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

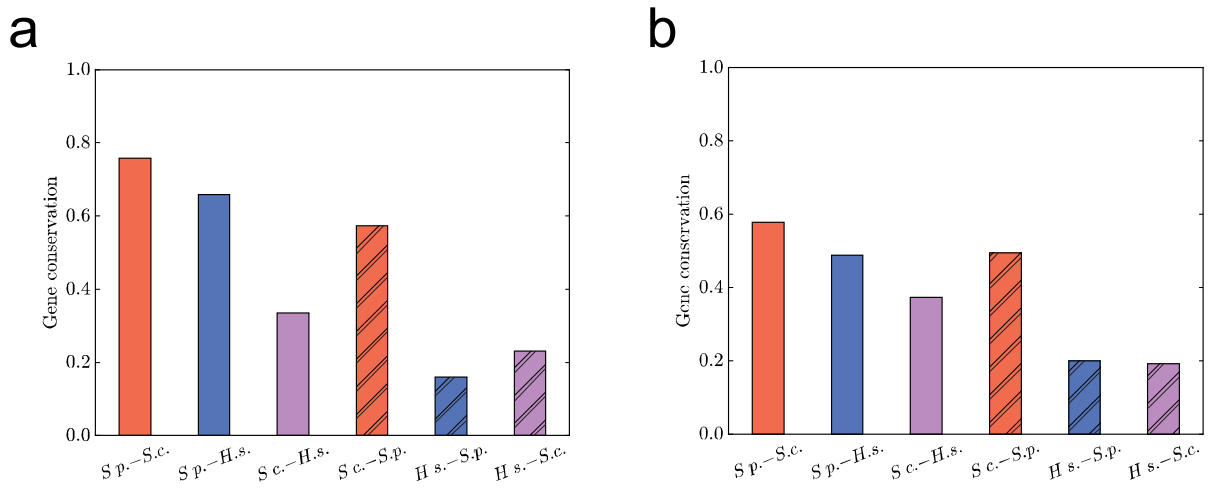

Figure A.2 Pairwise comparisons of the conservation of all coding genes between reference-query species pairs. (a) Orthologs are defined by PomBase (McDowall et al., 2015) and Saccharomyces Genome Database (SGD) (Cherry et al., 2012). (b) Orthologs are defined by InParanoid (Sonnhammer and Ostlund, 2015).

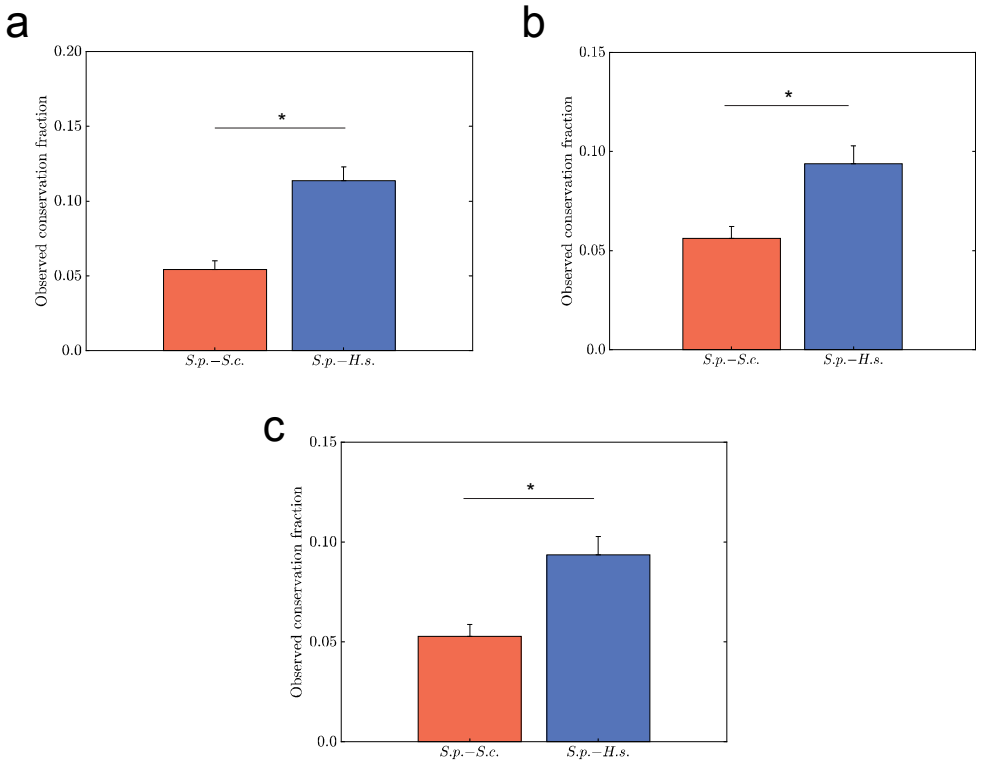

Figure A.3 (a) Observed interaction conservation between reference-query species pairs using co-crystal structures for *S. cerevisiae* and human. (b-c) Observed interaction conservation between reference-query species pairs using large-scale AP/MS datasets for S. cerevisiae and human. For both panels, the human AP/MS dataset used is from (Huttlin et al., 2015). (b) The *S. cerevisiae* AP/MS dataset is from (Gavin et al., 2006). (c) The *S. cerevisiae* AP/MS dataset is from (Krogan et al., 2006). Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

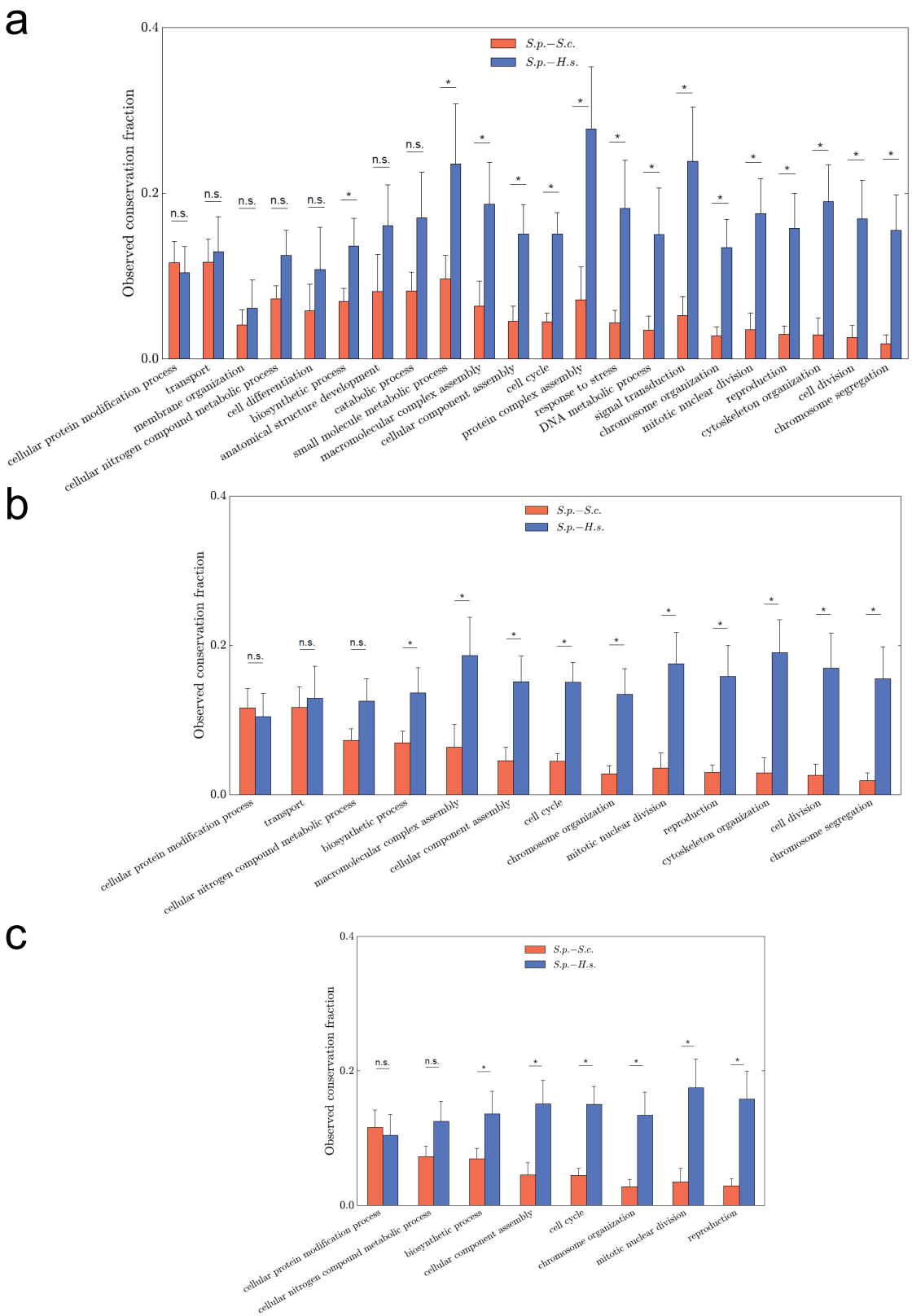

Figure A.4 Observed conservation fractions of *S. pombe* interactions in *S. cerevisiae* and human in different GO Slim biological process categories with at least (a) 30, (b) 50, and (c) 75 interactions. Data are shown as measurements + SE. * denotes significant (P<0.05); n.s. denotes not significant.

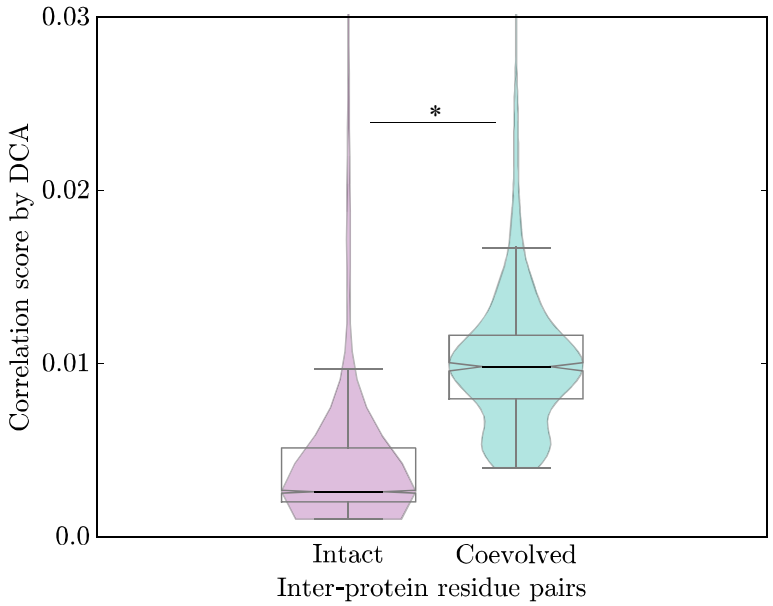

Figure A.5 Co-evolution analysis of proteins involved in intact or co-evolved interactions using DCA. * denotes significant (P<0.05).

A.16 REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R.*, et al.* (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res *40*, D700-705.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972-976.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B.*, et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631-636.

Hubbard, S. (1996). NACCESS v. 2.1.1 [Computer program]. http://wwwbioinfmanchesteracuk/naccess/.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K.*, et al.* (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell *162*, 425-440.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S.*, et al.* (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol *5*, R7.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P.*, et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637-643.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Matsuyama, A., Arai, R., Yashiroda, Y., Shirai, A., Kamata, A., Sekido, S., Kobayashi, Y., Hashimoto, A., Hamamoto, M., Hiraoka, Y.*, et al.* (2006). ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol *24*, 841-847.

McDowall, M.D., Harris, M.A., Lock, A., Rutherford, K., Staines, D.M., Bahler, J., Kersey, P.J., Oliver, S.G., and Wood, V. (2015). PomBase 2015: updates to the fission yeast database. Nucleic Acids Res *43*, D656-661.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A *108*, E1293-1301.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods *9*, 471-472.

Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bahler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. Nat Genet *36*, 809-817.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

Sonnhammer, E.L., and Ostlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res *43*, D234-239.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yu, H., Jansen, R., Stolovitzky, G., and Gerstein, M. (2007). Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. Bioinformatics *23*, 2163-2173.

APPENDIX B

Supplementary Information for:

mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome

B.1 Sources of the structural proteome

In order to build our repository of protein structures and models, we curated experimentally-determined crystal structures from the PDB (Berman et al., 2000) and homology models from ModBase (Pieper et al., 2011) by searching for Swiss-Prot (UniProt-Consortium, 2015) tagged structures or chains in both (see Methods). Over the past 20 years, the coverage of these databases has increased dramatically, enabling clustering on the scale of the whole proteome (Figure B.1).

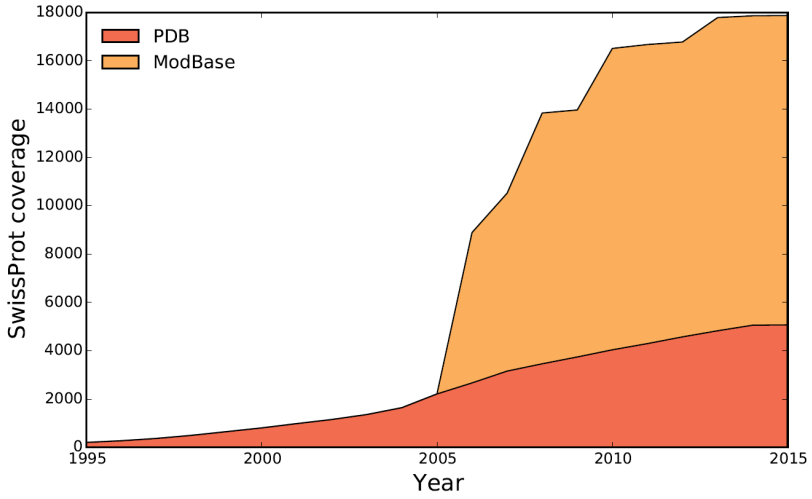

Figure B.1 The growth of the two sources of protein structures for mutation3D. By April 2015, the PDB and ModBase together accounted for ~ 88% of the verified human proteome (SwissProt) as measured by the number of unique entries in ModBase and PDB matching unique SwissProt IDs (for verified, canonical isoforms) divided by the total number of SwissProt proteins in UniProt. At the time of this calculation, there were 5,068 PDB entries representing unique SwissProt proteins (≥ 20 residues), 12,809 ModBase entries matching unique SwissProt proteins (≥ 20 residues and MPQS ≥ 0.5) not in the PDB, and 20,204 total SwissProt proteins catalogued by UniProt.

B.2 Model filter categories

mutation3D supplements its crystal structure coverage of the proteome with homology models derived from ModBase (Pieper et al., 2011). ModBase provides many quality metrics to determine the accuracy of each of their models, several of which we have included for filtering purposes. The user may select a subset of models by setting these parameters on the advanced page. All parameters are set to their most relaxed levels, except for *MPQS*, as this parameter encompasses all other parameters, thereby allowing some leniency in the others (i.e. one low quality parameter can be compensated for by several high quality parameters). ModBase states that a model should be considered to have a reliable fold assignment if any one of parameters 3-5 (below) fall within an acceptable range. Thus, setting all parameters to their suggested thresholds may inappropriately and unnecessarily remove certain models from further consideration. However, individual users may still choose to filter on different or additional parameters by setting any combination of the parameters below.

(1) *Protein Coverage*: The fraction of the full length UniProt (UniProt-Consortium, 2015) protein covered by the model, irrespective of the identity or accuracy of the 3D amino acid positions in the model. There is no suggested cutoff for *Protein Coverage* as it can vary based upon the user's specific requirements.

(2) *Sequence Identity*: The fraction of a model's amino acid sequence that is identical to the UniProt amino acid sequence, on a scale of 0 to 1. Note that this measure is independent of the *Protein Coverage* (i.e. *Sequence Identity* can still equal 1 if all of the amino acids included in the model are identical to amino acids in the covered region of the protein irrespective of how large this region is). There is no suggested cutoff for *Sequence Identity* as it can vary based upon the user's specific requirements.

(3) *e-value*: The significance of the alignment between the template PDB (Berman et al., 2000) structure sequence and the target UniProt sequence as reported by NCBI's PSI-BLAST program (Altschul et al., 1997) or similar alignment score calculated by ModPipe (Pieper et al., 2011). The ModBase-suggested quality cutoff using *e-value* alone is *e-value* < 10^{-4}.

(4) *Discrete Optimized Protein Energy (DOPE) score*: Also known as *zDOPE*, it is a derived atomic distance-dependent statistical potential calculated by ModBase from a sample of native structures. The ModBase publication describes this measure in great detail (Pieper et al., 2011). Lower *zDOPE* scores indicate a more accurate model. The ModBase-suggested quality cutoff using *zDOPE* score alone is *zDope* < 0.

(5) *ModPipe Quality Score (MPQS)*: A composite score calculated by ModBase comprising several measures including measures 1-4. Since this score incorporates all previous scores, mutation3D uses the ModBase-suggested threshold for a high quality model of *MPQS* ≥ 1.1 by default.

B.3 Clustering parameters

These parameters define the properties of an acceptable amino acid substitution cluster in mutation3D. Suggested values are pre-set in both the standard query interface and the advanced page, but the user may choose to change the parameters from the advanced page.

(1) *Minimum Number of Substitutions*: The minimum number of amino acid substitutions required to form a cluster. This refers to the absolute number of substitutions, irrespective of whether they exist at the same amino acid position or arise from the same underlying nucleotide mutation across multiple clinical samples. By default, the *Minimum Number of Substitutions* = 3.

(2) *Minimum Number of Unique Substitutions*: The minimum number of amino acid substitutions existing at unique amino acid indices required to form a cluster. In this sense, multiple mutations within one codon are counted as a single unique substitution, irrespective of whether or not they give rise to different mutant amino acid residues. By default, the *Minimum Number of Unique Substitutions* = 2.

(3) *Maximum Cluster Diameter*: The maximum allowable distance in Angstroms between any two amino acid substitutions in a cluster. This is the same as the CL-distance in Complete Linkage Clustering and can be thought of as the maximum allowable diameter of a sphere containing all points in a cluster. By default, the *Maximum Cluster Diameter* = 15 Å.

(4) *Minimum Linear Separation*: A post-clustering filter parameter to remove clusters that do not span a specified distance in amino acid index space (i.e. the number of amino acids separating two residues in a linear chain). Users interested in identifying examples of clusters that would only be observable in 3D space should set this parameter higher. By default, *Minimum Linear Separation* = 0.

B.4 Amino acid substitution patterns in the Ras GTPase protein family

To assess the plausibility of the postulate that driver mutations, because of their functional similarity, could occur as clustered amino acid substitutions in the same protein, we considered the canonical Ras family of cancer genes (*KRAS, NRAS, HRAS*) and their corresponding protein products, each 189 amino acids in length. According to COSMIC (Forbes et al., 2011), 99% of all somatic missense mutations in these genes occur in codons 12, 13 and 61 (Table B.3). That these mutations have been noted at such high frequencies in tumor tissues strongly supports the view that they are drivers of tumorigenesis rather than passengers (Stratton et al., 2009).

It is highly likely that the functional mechanisms by which missense mutations in codons 12 and 13 confer their tumorigenic phenotypes are closely related because of the assured proximity of the juxtaposed amino acids within the tertiary structures of their respective proteins. However, the mechanism by which missense mutations in codon 61 exert their tumorigenic effects is less clear, given its position on the protein backbone. In cases such as these, and systematically across genome-wide cancer mutation datasets, efforts to discern functionality shared by linearly remote but spatially clustered missense mutations must be made at a protein structural level using crystal structures and models.

In the case of the RAS-family protein products, there are many available crystal structures and models. In Figure 5.1b, we highlight the locations of the three known driver mutations in KRAS and show that they form a tight cluster within the crystal structure. This spatial relationship was revealed by mutation3D through clustering in many primary sequencing studies available through COSMIC that sequenced Ras-family genes, and suggests that the three substitutions in Ras-family proteins act in a mechanistically similar fashion so as to confer the tumorigenic phenotype. This conclusion is not of course novel since it has been shown many times before that the mechanisms by which these three mutations exert their effects in the RAS proteins are closely related and that the cancer phenotype is likely to be attributable to the constitutive binding of GTP (Pylayeva-Gupta et al., 2011). Such an example does however illustrate how mutation3D could in principle be used to explore previously unknown disease etiologies and potentially inform treatment regimens.

B.5 Oncogenes vs. tumor suppressors

We explored the ability of mutation3D to detect oncogenes and tumor suppressors. Since oncogenes tend to act via gain-of-function and tumor suppressors tend to act via loss-of-function, we

considered it likely that clusters would be found more frequently in oncogenes than in tumor suppressors, as gain of function may require more specific mutations than loss of function (Hanahan and Weinberg, 2000). However, we also considered that tumor suppressors may be more likely to contain missense mutations in the first place, while oncogenes are known to often act through somatic copy number alterations (Beroukhim et al., 2010). Overall we find that mutation3D does not preferentially detect either class of cancer gene, demonstrating robustness to detect both oncogenes and tumor suppressors (Figure B.2).

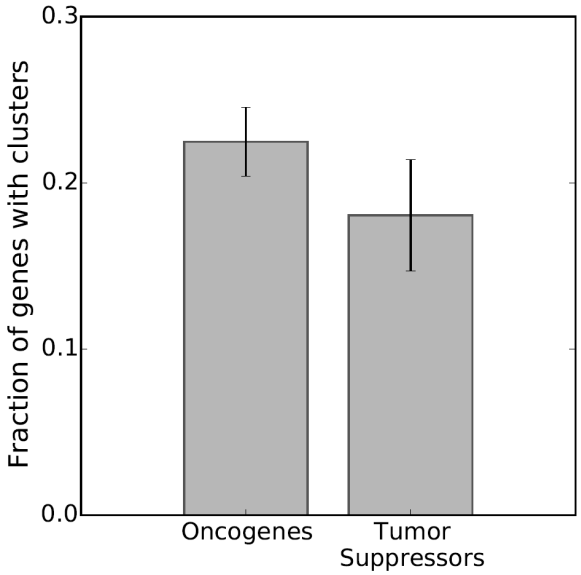

Figure B.2 The Cancer Gene Census annotates many genes as oncogenes or tumor suppressors. Plotted are the fractions genes in each set that have been detected by mutation3D with at least one cluster in COSMIC WGS studies.

B.6 Reduction-to-1D clustering methods

Two recent reduction-to-1D clustering methods, iPAC (Ryslik et al., 2013) and GraphPAC (Ryslik et al., 2014), take into account the 3D structure of proteins in order to identify non-random somatic mutations in cancer. The key difference between mutation3D and these methods is that mutation3D

performs 3D clustering by using the coordinates of α-carbons directly in protein models. Any cluster found by mutation3D therefore exists by definition in 3D space, and visualization of clusters using mutation3D's web interface demonstrates this. On the other hand, while reduction-to-1D methods make use of the 3D coordinate information in protein models, they first reduce the number of dimensions from 3 to 1 in order to use algorithms designed for 1D clustering, such as Non-Random Mutational Clustering (Ye et al., 2010) (NMC).

While clustering performed by NMC in 1D may be accurate given the projected 1D coordinates, the projected coordinates themselves will not retain all of the information of the original 3D coordinates. Although such reduction-to-1D methods attempt to minimize loss of information due to dimensionality reduction, they cannot eliminate it. Here, this loss of information can lead to both false positives and false negatives for the identification of clusters in the original 3D space. To illustrate this, we have shown in Figure B.3 how dimensionality reduction using Multidimensional scaling (Borg, 1997) (MDS) to transform 3D coordinates into 1D can lead to both false positive and false negative 3D clusters in KRAS.

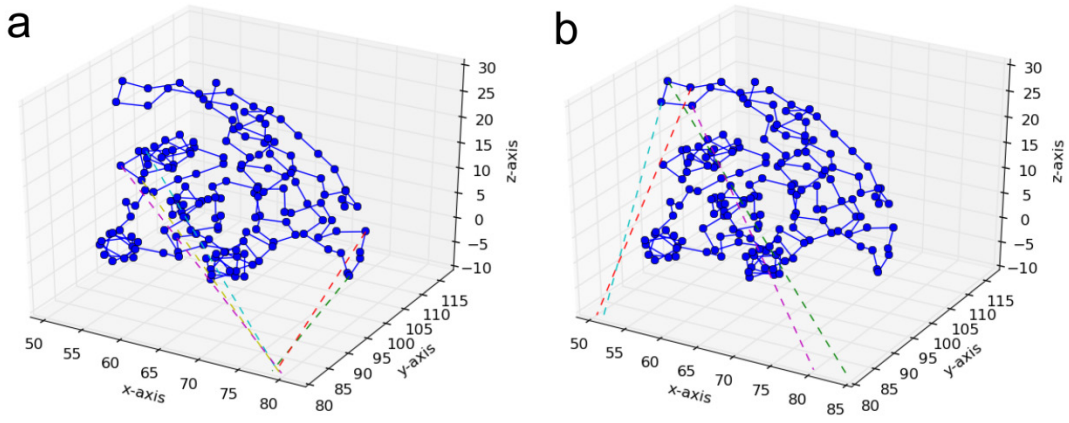

Figure B.3 α-carbons in KRAS are plotted in 3D space according to the coordinates provided by the PDB (3GFT). In each figure panel, selected projections from 3D to 1D coordinates using MDS are shown scaled between 50 and 85 on the x-axis. (a) α-carbons in two distinct areas of the 3D structure are projected very near in 1D space. While any of these could potentially be clustered based on the 1D projections, not all of the initial α-carbon positions are close in 3D space, resulting in false positive identification of clusters. (b) Four α-carbons close in 3D space are very far apart when projected into 1D using MDS, resulting in false negatives.

B.7 Statistical bootstrapping model

Clusters in mutation3D are found in protein structures and models using an implementation of complete-linkage clustering. These clusters exist by virtue of the spatial arrangement of amino acid substitutions. However, in order to assess whether these clusters represent significant findings or simply arise by chance we must employ a model of statistical significance.

There are many methods (Kan et al., 2010; Lawrence et al., 2013; Sjöblom et al., 2006) already available to address the hypermutation hypothesis—that genes with a number of mutations far above expectation given background mutations rates are likely to be involved in cancer. We make a distinction from these methods in order to test whether the spatial arrangement of mutations on the protein backbone alone is significant. In other words, do amino acid substitutions occur in non-random proximity given the contours of the protein backbone?

In order to test this hypothesis, we must perform randomized iterations of amino acid substitution placement to produce a background distribution (see Materials and Methods), as each protein model and structure will have a different null expectation. For instance, a highly bunched structure will be very likely to produce clusters if all residues are with a short distance from each other. On the other hand, a nearly linear structure, with residues maximally far apart from each other, will be less likely to produce clusters given the same number of randomly chosen substitution positions.

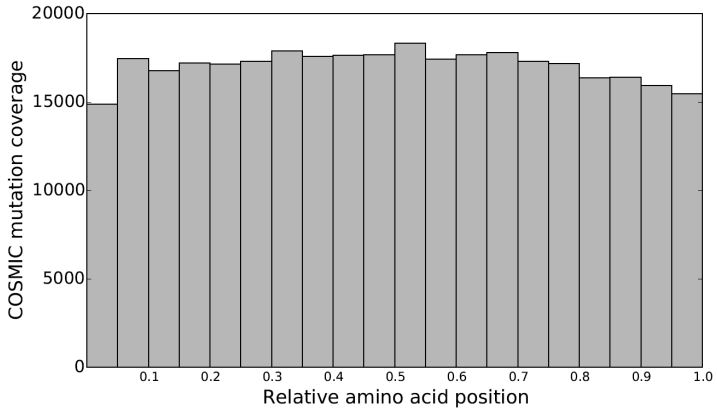

Figure B.4 The relative positions of all amino acid substitutions in 175 whole genome studies in COSMIC. Relative position = (amino acid index)/(protein length).

There is a substantial body of literature (Ramsey et al., 2011; Toth-Petroczy and Tawfik, 2011; Tusche et al., 2012; Zhou et al., 2008) describing analyses of positive selection rates suggesting that amino acids on the surface of a protein are more likely to be mutated than those that are buried. When this is the case, a statistical model designed to identify clusters should adjust for the difference in these substitution rates to avoid finding spurious clusters on protein surfaces. However, we note that amino acids substitutions in WGS cancer screens do not preferentially occur near the C or N termini of proteins (Figure B.5), which should be more likely to be exposed in protein structures (this litmus test has been used in these evolutionary analyses to propose the need to adjust for the differences in observed rates of surface and buried residues). Therefore, since we do not observe a difference in substitution rates between buried and surface residues in cancer, and because mutation3D has been designed for a more general use case to detect clusters in any dataset, mutation3D does not treat buried and surface residues differently in its iterative bootstrapping model.

Nucleotide Change	Amino Acid Substitution
c.254T>G	M85R
c.380T>G	M127R
c.1094G>A	R365Q
c.1108G>A	V370M
c.1123C>T	R375C
c.1124G>A	R375H
c.1232A>G	N411S
c.1303C>T	R435C
c.1310G>A	C437Y

Table B.1 The inherited disease associated amino acid substitutions from HGMD shown in Figure 5.1a for aromatase (*CYP19A1* gene, transcript: NM_031226.2). Nucleotides are indexed in coding sequences, using the A of the ATG translation initiation start site as nucleotide 1.

dbSNP ID	Nucleotide Change	Amino Acid Substitution	ESP 6500 Allele Frequency
rs28757184	c.602C>T	T201M	0.04446
rs700519	c.790C>T	R264C	0.077374

Table B.2 The SNPs and their associated amino acid substitutions as shown in Figure 5.1a of aromatase (*CYP19A1* gene, transcript: NM_031226.2). Both SNPs are predicted to be benign by PolyPhen-2. Nucleotides are indexed in coding sequences, using the A of the ATG translation initiation start site as nucleotide 1.

		Protein Product		
		HRAS (P01112)	KRAS (P01116)	NRAS (P01111)
Codon	12	419 (G>V)	23,742 (G>D)	746 (G>D)
	13	118 (G>R)	4,101 (G>D)	373 (G>D)
	61	367 (Q>R)	367 (Q>H)	1,863 (Q>R)
	All	940	28,444	3,036

Table B.3 The distribution of amino acid substitutions in Ras GTPase codons reported in COSMIC v67. Given in parentheses are the most common substitutions for each codon in each protein.

Figure	Input Data	Unique amino acid indices	Amino acids	ModBase MPQS	Maximum cluster diameter (Å)	*P*-value	Output Measurement
5.3a/b	HGMD and ESP	≥ 2	≥ 3	≥ 1.1	15	n/a	Fraction of substitutions clustered
5.3c	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	5, 10, 15, 20, 25	n/a	Fraction of known cancer genes
5.3d	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	25	10^{0}, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}	Fraction of known cancer genes
5.3e	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	5, 15, 25	n/a	PolyPhen-2 scores of clustered mutations
5.3f	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	5, 15, 25	n/a	Frequency of clustered mutations in COSMIC
5.4a	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	15	≤ 0.01	Top genes predicted by mutation3D

Table B.4 Computational parameters for Figure 5.3. Each panel 5.3a-f is designed to test only one parameter at a time (i.e. protein model set, *P*-value, or cluster diameter). Figure 5.5 combines recommended parameters to predict cancer genes.

Note: Header 'Unique amino acid indices' refers to the minimum number of unique amino acid positions needed to be mutated to constitute a cluster. Header 'Amino acids' refers to the minimum total number of amino acids (including at the same indices) needed to be mutated to constitute a cluster. Default values for these parameters are used throughout the study (see Section B.3).

Gene	UniProt	# Studies	CGC?	Mutsig?	Most common		Most Significant Cluster			
					Tissue	Mutations	Tissue	Cluster	PMID	*P*-value
TP53	P04637	101	Y	Y	Central Nervous System(12)	248(51), 273(47), 175(45), 245(31), 179(26)	Lung	R175L(6), M246I(5), R248L(5)	23856246	<6.67E-05
KRAS	P01116	33	Y	Y	Large Intestine(10)	12(30), 13(21), 61(17), 146(13), 117(5)	Pancreas	G12V(80), G12D(74), G12R(24), Q61H(10), G12C(8), Q61R(2), G12S(2), Q61K(2), G13D(2)	25855536	<6.67E-05
PIK3CA	P42336	24	Y	Y	Large Intestine(4)	545(16), 542(15), 546(10), 1047(8), 420(5)	Ovary	E545K(4), E542K(2), Q546K(1), E542V(1), E545V(1), E545A(1)	20826764	<6.67E-05
FBXW7	Q969H0	16	Y	Y	Large Intestine(6)	505(11), 465(10), 479(4), 626(2), 423(2)	Large Intestine	V464M(5), R465H(5)	24599305	<6.67E-05
NRAS	P01111	10	Y	Y	Haematopoietic And Lymphoid Tissue(5)	12(10), 13(8), 61(7), 68(1), 66(1)	Haematopoietic And Lymphoid Tissue	G12S(3), G13D(2), G12V(1), G13R(1), Q61R(1), Q61K(1), G12A(1)	25381062	<6.67E-05
HRAS	P01112	9	Y	Y	Upper Aerodigestive Tract(5)	61(9), 12(8), 13(7), 58(1), 117(1)	Skin	Q61L(4), T58I(2), G12S(2), Q61K(2), G13D(2)	25303977	<6.67E-05
SMAD4	Q13485	9	Y	Y	Large Intestine(5)	361(7), 356(6), 351(4), 386(4), 524(2)	Oesophagus	G386D(3), D351H(1)	23525077	<6.67E-05
CDKN2A	Q8N726	8	Y	Y	Large Intestine(2)	94(5), 102(4), 97(3), 98(2), 107(1)	Oesophagus	H107Q(4), R98L(4), G102V(1), P94L(1)	25839328	<6.67E-05
FRG1B	Q9BZ01	8	N	N	Large Intestine(2)	80(3), 101(3), 50(2), 60(2), 88(2)	Skin	Y46H(4), L50P(4), M11V(2)	24265154	<6.67E-05
GRIK2	Q13002	8	N	N	Skin(3)	830(3), 763(2), 216(1), 213(1), 764(1)	Lung	V526L(2), E524D(2)	22941189	<6.67E-05
NRXN1	Q9ULB1	8	N	N	Large Intestine(3)	1216(2), 1077(2), 1151(1), 856(1), 1167(1)	Large Intestine	A660T(4), C643Y(4)	24211491	<6.67E-05
EGFR	P00533	7	Y	Y	Lung(3)	858(3), 62(2), 289(2), 719(2), 768(1)	Lung	L858R(11), L833V(1), R889G(1)	25189529	<6.67E-05
GRIA2	P42262	7	N	N	Large Intestine(3)	692(1), 752(1), 696(1), 718(1), 335(1)	Oesophagus	T501N(2), G752C(2)	25839328	<6.67E-05
BRAF	P15056	6	Y	Y	Lung(2)	469(5), 600(4), 601(3), 597(2), 466(2)	Urinary Tract	G596R(1), F595L(1), D594G(1), G469A(1)	24121792	<6.67E-05
CDKN2A	P42771	6	Y	Y	Oesophagus(2)	83(3), 114(2), 84(2), 102(2), 21(1)	Skin	P114L(6), E88K(3), H83R(2)	22817889	<6.67E-05
EPHA3	P29320	6	N	N	Large Intestine(3)	48(1), 83(1), 75(1), 94(1), 357(1)	Oesophagus	N79I(2), D75G(2)	23525077	0.00085
GRM8	O00222	6	N	N	Large Intestine(2)	38(1), 212(1), 465(1), 261(1), 53(1)	Liver	G465E(3), D407N(3)	25822088	<6.67E-05
PLXNA4	Q9HCM2	6	N	N	Large Intestine(4)	333(2), 591(1), 459(1), 318(1), 1326(1)	Large Intestine	A437T(4), R459W(2)	25344691	0.00415
DPP10	Q8N608	5	N	N	Large Intestine(2)	152(2), 144(1), 140(1), 603(1), 571(1)	Skin	E220K(2), S152L(2)	21984974	<6.67E-05
EPHB1	P54762	5	N	N	Stomach(1)	846(1), 691(1), 616(1), 190(1), 117(1)	Ovary	T117N(2), A176T(2)	21720365	<6.67E-05
FCAR	P24071	5	N	N	Large Intestine(4)	178(2), 128(2), 155(1), 195(1), 51(1)	Large Intestine	L128P(2), F178L(2)	23856246	<6.67E-05
ITGB8	P26012	5	N	N	Large Intestine(2)	146(1), 339(1), 140(1), 330(1), 333(1)	Upper Aerodigestive Tract	C481Y(2), N474K(2)	25275298	<6.67E-05
KCNMA1	Q12791	5	N	N	Large Intestine(2)	630(1), 604(1), 1096(1), 909(1), 905(1)	Skin	R865C(3), D1096N(3)	22842228	<6.67E-05
LCK	P06239	5	Y	N	Large Intestine(3)	458(2), 484(2), 151(1), 197(1), 291(1)	Haematopoietic And Lymphoid Tissue	A289D(2), A396V(2)	23856246	0.000133
MGAM	O43451	5	N	N	Skin(2)	246(2), 115(1), 123(1), 1045(1), 790(1)	Ovary	K281R(2), V263M(2)	21720365	<6.67E-05
MYH13	Q9UKX3	5	N	N	Lung(1)	590(1), 58(1), 712(1), 108(1), 109(1)	Oesophagus	D58E(2), P79L(2)	25151357	<6.67E-05
NTRK3	Q16288	5	Y	N	Skin(2)	584(1), 583(1), 746(1), 747(1), 382(1)	Stomach	F747V(2), K746T(2)	24816253	<6.67E-05
OR1L8	Q8NGR8	5	N	N	Large Intestine(2)	201(1), 200(1), 127(1), 247(1), 123(1)	Skin	H6Y(2), N5S(2)	25303977	<6.67E-05
PDE10A	Q9Y233	5	N	N	Large Intestine(2)	752(1), 338(1), 719(1), 704(1), 405(1)	Liver	M704I(2), A722V(2)	25822088	<6.67E-05
PIK3R1	P27986	5	Y	Y	Haematopoietic And Lymphoid Tissue(2)	567(2), 573(1), 464(1), 452(1), 564(1)	Haematopoietic And Lymphoid Tissue	L573P(3), N564K(3), K567E(3)	23143597	<6.67E-05
SYK	P43405	5	Y	N	Large Intestine(2)	33(1), 330(1), 29(1), 52(1), 428(1)	Lung	K104R(1), K105N(1), K105E(1)	22980975	<6.67E-05
ABCB5	Q2M3G0	4	N	N	Soft Tissue(2)	678(1), 581(1), 560(1), 596(1), 680(1)	Pancreas	K560E(3), M525I(3)	25855536	<6.67E-05
ACO1	P21399	4	N	Y	Large Intestine(2)	378(2), 263(2), 448(1), 41(1), 61(1)	Large Intestine	T263I(2), D378N(2)	23856246	<6.67E-05
ADSL	P30566	4	N	N	Large Intestine(4)	300(2), 296(2), 24(1), 77(1), 354(1)	Large Intestine	R300H(6), A291V(2), R296Q(2)	24755471	<6.67E-05
CASP8	Q14790	4	Y	Y	Large Intestine(3)	228(2), 232(2), 233(2), 236(1), 237(1)	Large Intestine	M228I(3), P232H(3)	23856246	<6.67E-05
CHAT	P28329	4	N	N	Large Intestine(1)	217(1), 618(1), 174(1), 669(1), 610(1)	Liver	A217V(2), C669S(2)	25822088	<6.67E-05
CLVS2	Q5SYC1	4	N	N	Large Intestine(1)	135(1), 277(1), 99(1), 109(1), 265(1)	Liver	A109T(2), L265M(2)	25822088	<6.67E-05
CTBP2	P56545	4	N	N	Large Intestine(2)	157(1), 336(1), 82(1), 64(1), 341(1)	Haematopoietic And Lymphoid Tissue	R157W(4), T160K(4)	24970810	0.00017
CTNNA2	P26232	4	N	N	Stomach(2)	753(1), 323(1), 893(1), 756(1), 754(1)	Stomach	D241E(2), R240P(2)	25042771	<6.67E-05

CTNNA3	Q9UI47	4	N	N	Large Intestine(3)	214(1), 842(1), 834(1), 415(1), 158(1)	Large Intestine	R842L(2), R834Q(2)	24211491	<6.67E-05
EPHA4	P54764	4	N	N	Oesophagus(1)	775(1), 604(1), 771(1), 154(1), 773(1)	Lung	D773N(2), D604Y(2)	22980975	<6.67E-05
ERCC2	P18074	4	Y	Y	Urinary Tract(1)	312(1), 44(1), 42(1), 463(1), 181(1)	Bone	K181T(2), R185Q(2)	25186949	<6.67E-05
FN1	P02751	4	N	N	Large Intestine(2)	414(1), 447(1), 1658(1), 2252(1), 1861(1)	Oesophagus	R2266H(2), H2252R(2)	25839328	<6.67E-05
GCK	P35557	4	N	N	Large Intestine(2)	147(1), 21(1), 156(1), 63(1), 304(1)	Skin	F330Y(2), G328R(2)	22842228	<6.67E-05
GNAS	Q5JWF2	4	Y	N	Upper Aerodigestive Tract(1)	844(2), 740(1), 869(1), 961(1), 874(1)	Upper Aerodigestive Tract	R844C(2), N740S(2)	21798893	<6.67E-05
GNAS	P63092	4	Y	N	Upper Aerodigestive Tract(1)	201(2), 318(1), 338(1), 227(1), 226(1)	Peritoneum	R201C(22), R201H(8), R201L(2), Q227H(2)	24944587	<6.67E-05
GRIA4	P48058	4	N	N	Large Intestine(2)	627(1), 634(1), 561(1), 45(1), 342(1)	Skin	G367R(3), G342R(3)	25303977	0.00302
GRIK1	P39086	4	N	N	Large Intestine(2)	797(1), 857(1), 610(1), 218(1), 122(1)	Stomach	V128G(3), I122M(3)	24816253	8.99E-05
GRIK3	Q13003	4	N	N	Large Intestine(1)	450(1), 301(1), 440(1), 447(1), 455(1)	Large Intestine	T455M(4), R450Q(2)	25344691	0.000321
GRM1	Q13255	4	N	N	Large Intestine(2)	309(1), 44(1), 369(1), 43(1), 460(1)	Large Intestine	V309M(2), R275H(2), E311K(2)	25344691	0.000234
GRM3	Q14832	4	N	N	Large Intestine(2)	59(1), 49(1), 46(1), 460(1), 306(1)	Skin	F187L(2), S182L(2), G460E(1)	25303977	0.00302
HK3	P52790	4	N	N	Large Intestine(2)	214(1), 215(1), 772(1), 676(1), 777(1)	Haematopoietic And Lymphoid Tissue	P676S(2), Y673C(2)	23292937	<6.67E-05
HLA-B	P30486	4	N	Y	Upper Aerodigestive Tract(1)	201(1), 202(1), 140(1), 119(1), 107(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	Q29836	4	N	Y	Upper Aerodigestive Tract(1)	91(2), 201(1), 202(1), 155(1), 140(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	P01889	4	N	Y	Upper Aerodigestive Tract(1)	140(2), 91(2), 155(1), 137(1), 69(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	Q31610	4	N	Y	Upper Aerodigestive Tract(1)	91(2), 201(1), 202(1), 155(1), 140(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	Q31612	4	N	Y	Upper Aerodigestive Tract(1)	91(2), 155(1), 195(1), 197(1), 176(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-DRB1	P01911	4	N	N	Upper Aerodigestive Tract(1)	135(2), 195(2), 127(2), 133(2), 197(1)	Pancreas	K127Q(5), S133A(2)	23912084	<6.67E-05
HLA-DRB1	Q9GIY3	4	N	N	Upper Aerodigestive Tract(1)	195(2), 11(1), 60(1), 12(1), 59(1)	Pancreas	K127Q(5), S133A(2)	23912084	<6.67E-05
HS3ST4	Q9Y661	4	N	N	Large Intestine(2)	200(1), 313(1), 301(1), 218(1), 411(1)	Lung	T299M(2), K411N(2), T301I(2)	22980975	<6.67E-05
HTR4	Q13639	4	N	N	Large Intestine(2)	302(1), 306(1), 77(1), 183(1), 242(1)	Large Intestine	Y302C(3), G306R(2)	22810696	<6.67E-05
INSRR	P14616	4	N	N	Skin(2)	995(1), 1164(1), 1138(1), 1160(1), 1171(1)	Skin	S1163F(2), T1171P(2)	22842228	<6.67E-05
IPO11	Q9UI26	4	N	N	Large Intestine(2)	797(1), 587(1), 835(1), 337(1), 589(1)	Bone	R835Q(2), R797Q(2)	25186949	<6.67E-05
MCM7	P33993	4	N	N	Large Intestine(3)	415(1), 611(1), 455(1), 445(1), 356(1)	Large Intestine	L356F(2), R611H(2)	22895193	<6.67E-05
MSRB3	Q8IXL7	4	N	N	Large Intestine(2)	184(1), 53(1), 77(1), 63(1), 71(1)	Large Intestine	S161L(2), H77D(2), F71I(2)	22810696	<6.67E-05
NCK2	O43639	4	N	N	Large Intestine(2)	39(1), 205(1), 213(1), 40(1), 273(1)	Large Intestine	D257E(6), I225T(2)	24755471	<6.67E-05
NRXN1	P58400	4	N	N	Skin(2)	181(2), 109(1), 115(1), 132(1), 116(1)	Large Intestine	G225D(4), R132Q(4)	22895193	<6.67E-05
OPRM1	P35372	4	N	N	Large Intestine(2)	347(1), 322(1), 63(1), 46(1), 259(1)	Skin	R369C(3), R347Q(3)	21984974	<6.67E-05
PDE1C	Q14123	4	N	N	Skin(2)	280(1), 468(1), 443(1), 454(1), 176(1)	Lung	R522T(2), K524T(2)	22980975	0.000606
PFKP	Q01813	4	N	N	Large Intestine(2)	467(1), 463(1), 60(1), 177(1), 130(1)	Lung	W463C(2), G467A(2)	22980975	<6.67E-05
PLCB1	Q9NQ66	4	N	N	Skin(2)	720(2), 740(1), 569(1), 768(1), 696(1)	Large Intestine	I767N(3), E302G(2)	24755471	0.00121
POLD1	P28340	4	N	N	Oesophagus(1)	101(1), 339(1), 697(1), 455(1), 306(1)	Lung	I101T(2), V455L(2), G422C(2)	22980975	<6.67E-05
POT1	Q9NUX5	4	Y	N	Haematopoietic And Lymphoid Tissue(2)	105(2), 36(1), 77(1), 116(1), 137(1)	Liver	T105M(2), P116S(2)	25822088	<6.67E-05
PRSS3	P35030	4	N	N	Skin(2)	146(2), 265(2), 148(2), 270(2), 179(2)	Haematopoietic And Lymphoid Tissue	T86I(4), K216Q(4)	24970810	<6.67E-05
PSMB11	A5LHX3	4	N	N	Large Intestine(3)	206(2), 215(1), 65(1), 169(1), 106(1)	Large Intestine	R169C(3), F161V(3)	25344691	<6.67E-05
PTPRD	P23468	4	N	N	Skin(2)	1898(2), 1647(1), 1828(1), 49(1), 1843(1)	Oesophagus	L1377V(4), N1388K(4), A1387T(4)	23525077	0.00662
SEMA3C	Q99985	4	N	N	Large Intestine(3)	558(2), 522(2), 528(2), 526(1), 471(1)	Large Intestine	A522T(4), G558E(2)	24755471	<6.67E-05
SKIV2L	Q15477	4	N	N	Large Intestine(3)	752(1), 751(1), 769(1), 745(1), 754(1)	Large Intestine	L754F(1), I751M(1), L752P(1)	22810696	0.0002
SLC6A2	P23975	4	N	N	Large Intestine(2)	318(1), 140(1), 442(1), 562(1), 148(1)	Large Intestine	A145T(2), A562T(2)	22810696	<6.67E-05

SPAM1	P38567	4	N	N	Large Intestine(2)	207(1), 415(1), 132(1), 416(1), 130(1)	Kidney	T416P(2), F415L(2)	25401301	<6.67E-05
TGFBR2	P37173	4	N	Y	Large Intestine(3)	452(2), 454(2), 446(2), 524(2), 528(2)	Large Intestine	L452P(2), L454P(2)	23856246	<6.67E-05
TMPRSS3	P57727	4	N	N	Large Intestine(1)	400(1), 307(1), 404(1), 417(1), 326(1)	Skin	L307F(3), G417R(2)	25303977	<6.67E-05
TPO	P07202	4	N	N	Large Intestine(1)	153(1), 152(1), 335(1), 613(1), 461(1)	Lung	I613L(2), A640D(2)	22696596	<6.67E-05
TTN	Q8WZ42	4	N	N	Large Intestine(2)	32451(2), 32459(1), 2082(1), 32452(1), 2080(1)	Large Intestine	E32425G(12), K32459E(8), Y32452H(8)	24755471	0.00662

Table B.5 Genes whose protein products contain clusters of mutations in at least 4 COSMIC WGS studies. Clustering was performed using default parameters (see Section B.3), with a *P*-value cutoff of 0.01.

For each gene, reported are: the UniProt protein product, number of WGS studies in which clusters for the gene were identified, whether (Y) or not (N) the gene is in the Cancer Gene Census, whether (Y) or not (N) the gene was identified by MutSig, the primary tissue in which most clusters were found (and the number of studies in which clusters were found in that tissue), the top 5 most commonly mutated amino acid positions (and the number of studies in which mutations were found), and the most significant cluster found. The most significant cluster contains the following associated information: the tissue in which the cluster was found, the mutations in the cluster, the PMID of the publication reporting the mutations, and the *P*-value of the cluster.

Note: a *P*-value of <6.67E-05 indicates that bootstrapping produced no random mutation arrangements as tightly clustered as the observed data.

B.8 References

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389-3402.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Research *28*, 235-242.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M*., et al.* (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899-905.

Borg, I., Groenen, P. J. F. (1997). Modern multidimensional scaling : theory and applications.

Forbes, S., Bindal, N., Bamford, S., Cole, C., Kok, C., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A*., et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research *39*, 50.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.

Kan, Z., Jaiswal, B., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H., Yue, P., Haverty, P., Bourgon, R., Zheng, J*., et al.* (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. Nature *466*, 869-873.

Lawrence, M., Stojanov, P., Polak, P., Kryukov, G., Cibulskis, K., Sivachenko, A., Carter, S., Stewart, C., Mermel, C., Roberts, S*., et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Pieper, U., Webb, B., Barkan, D., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E., Pettersen, E., Huang, C*., et al.* (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research *39*, 74.

Pylayeva-Gupta, Y., Grabocka, E., and Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. Nature Reviews Cancer *11*, 761-774.

Ramsey, D.C., Scherrer, M.P., Zhou, T., and Wilke, C.O. (2011). The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics *188*, 479-488.

Ryslik, G.A., Cheng, Y., Cheung, K.-H.H., Modis, Y., and Zhao, H. (2013). Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics *14*, 190.

Ryslik, G.A., Cheng, Y., Cheung, K.-H.H., Modis, Y., and Zhao, H. (2014). A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics *15*, 86.

Sjöblom, T., Jones, S., Wood, L., Parsons, D., Lin, J., Barber, T., Mandelker, D., Leary, R., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.

Stratton, M., Campbell, P., and Futreal, P. (2009). The cancer genome. Nature *458*, 719-724.

Toth-Petroczy, A., and Tawfik, D.S. (2011). Slow protein evolutionary rates are dictated by surface-core association. Proceedings of the National Academy of Sciences of the United States of America *108*, 11151-11156.

Tusche, C., Steinbruck, L., and McHardy, A.C. (2012). Detecting patches of protein sites of influenza A viruses under positive selection. Mol Biol Evol *29*, 2063-2071.

UniProt-Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

Ye, J., Pavlicek, A., Lunney, E., Rejto, P., and Teng, C.-H. (2010). Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics *11*, 11.

Zhou, T., Enyeart, P.J., and Wilke, C.O. (2008). Detecting clusters of mutations. PloS one *3*, e3765.

APPENDIX C

Supplementary Information for:

A pan-interactome map of protein interaction interfaces

C.1 Interaction datasets

We compiled all high quality interactions available for *H. sapiens*, *D. melanogaster*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, *E. coli*, *S. pombe*, and *M. musculus* from HINT (Das and Yu, 2012). HINT employs a strict filtering strategy to ensure that interactions maintained in the high-quality sets are reliable. Those interactions with known interface residues based on available co-crystal structures in the Protein Data Bank (PDB) (Berman, 2000) were set aside for use in training and testing the classifier. Interactions without known interface residues comprise the set for which we make predictions.

For those interactions with known co-crystal structures in the PDB, we calculate interface residues for their specific binding partners. To identify UniProt protein sequences in the PDB, we use SIFTS (Velankar et al., 2013), which provides a mapping of PDB-indexed residues to UniProt-indexed residues (UniProt-Consortium, 2015). For each interaction and representative co-crystal structure, interface residues are calculated by assessing the change in solvent accessible surface area of the proteins in complex and apart using NACCESS (Lee and Richards, 1971). Any residue that is at the surface of a protein (≥15% exposed surface) and whose solvent accessible surface area (SASA) decreases by ≥1.0 $Å^2$ in complex is considered to be at the interface. We aggregate interface residues across all available structures in the PDB for a given interaction, wherein a residue is considered to be at the interface of the interaction if it has been calculated to be at the interface in one or more co-crystal structures of that interaction (all other residues are considered to be away from the interface). In building our final training and testing sets, we only consider interactions for

which aggregated co-crystal structures have combined to cover at least 50% of UniProt residues for both interacting proteins.

The training and testing sets each include a random selection of 400 interactions with known co-crystal structures, of which 200 are heterodimers and 200 are homodimers. To ensure an unbiased performance evaluation, we disallowed any homologous interactions (i.e. interactions whose structures could be used as templates for homology modeling) between the training and testing set. We also disallowed repeated proteins between the two sets to avoid simply reporting a remembered shared interface between a protein and multiple binding partners, thereby artificially elevating the performance of our classifier on the testing set.

C.2 Features

Biophysical Residue Properties: Seven per-residue biophysical sequence features are derived from ExPASy ProtScale (Artimo et al., 2012): (1) hydrophobicity, (2) polarity, (3) average buried area from protein folding, (4) molar fraction of accessible residues, (5) transmembrane tendency, (6) bulkiness, and (7) amino acid composition. Each residue in interacting proteins is assigned the value for that residue based on the published scales.

Evolutionary sequence conservation: We use PSI-BLAST (Altschul et al., 1997) to find homologs of each input protein across all proteins in the UniProt database. We retain the highest-scoring homolog by e-value for each organism in the PSI-BLAST output, with an e-value cutoff of 0.05. We then produce a multiple sequence alignment (MSA) between the original sequence and the retained UniProt sequences using Clustal Omega (Sievers et al., 2011). For both proteins in an interaction, we calculate the Jensen-Shannon divergence across all positions in the original protein sequence (Capra and Singh, 2007).

Co-evolution: Using the MSAs produced for the calculation of sequence conservation, we perform a co-evolutionary analysis of compensatory residue changes in interacting proteins. For each interaction, we search the MSAs of both proteins for homologs from the same organism. For species that have homologs for both interacting proteins, we concatenate the homologous sequences end-to-end. We then produce a new MSA containing concatenated sequences from all species for which there were homologs for both proteins. Using this MSA, we perform DCA (Morcos et al., 2014) and SCA (Lockless and Ranganathan, 1999) to evaluate the extent to which sequence positions between proteins are correlated over evolution. For DCA, we calculate both the direct information and mutual information scores. Because it is impossible to disentangle intra from inter-protein co-evolution for homodimers, we only perform co-evolution analyses for heterodimers.

Surface Residues: We curate both experimentally determined crystal structures from the PDB (Berman, 2000) and homology models from ModBase (Pieper et al., 2011) for each protein. For PDB structures, we calculate the solvent accessible surface area (SASA) using NACCESS (Lee and Richards, 1971) of all PDB chains which contain UniProt sequences, each chain in isolation. For ModBase models, we calculate SASA for all models with ModPipe Quality Score (MPQS) ≥ 1.1. For each protein, we average the raw values of SASA for each UniProt position covered by either a PDB structure or a ModBase model into a single feature.

Molecular Docking: Using the molecular subunits identified for the calculation of SASA from both the PDB and ModBase, we performed rigid-body molecular docking using zDOCK (Pierce et al., 2011). We performed docking for interactions where both proteins have a PDB structure covering ≥50% and 50 residues of each UniProt sequence or a ModBase model with the same sequence coverage and MPQS ≥1.1. In each case, docking was performed on the pair of structures with the highest overall UniProt residue coverage using zDOCK to produce 2,000 conformations of the

subunits in complex. Docking results are encoded as features by calculating the distance between each residue and the closest residue of the other subunit across the top 10 docking results.

C.3 Feature Engineering

Several strategies of transformation and aggregation of features were tested during cross-validation to determine which combinations were most predictive of interface residues.

Aggregation: Features that were collected from multiple sources were aggregated using a min, max, mean, or top strategy. SASA was collected across all available models, with each residue feature defined as either the maximum or mean SASA observed at that residue across all models. Docking features were aggregated as either the minimum, maximum, or mean distance from each residue to the opposing protein across the top 10 docked models. Co-evolution features were encoded for each residue as either the maximum, mean, or mean of the top 10 of the co-evolution scores with all residues in the interacting protein.

Normalization: Both raw feature values and normalized feature values were tested for many feature categories. Normalization was performed by calculating the z-score (number of standard deviations from the mean) of a residue feature in relation to all other such residue features in a given protein. When aggregation and normalization were performed together, aggregation preceded normalization.

C.4 Hyperparameter optimization with TPE

The tree-structured Parzen estimator approach (TPE) (Bergstra et al., 2011) is a Bayesian method for optimizing hyperparameters for machine learning algorithms. TPE models the probability distribution *p(x|y)* of hyperparameters given evaluated loss from a defined objective function, *L(x)*. We selected the following loss function to minimize based on classical hyperparameter inputs and residue window sizes:

$$L(\theta, w) = 1 - \min_{n \in \{1,2,3\}} \{AUROC_{\theta,w,n}\}$$

where x is comprised of θ, a set of hyperparameters, and w, a set of residue window sizes. The evaluation metric, $AUROC_n$, is the area under the roc curve for the n^{th} left-out evaluation fold in a three-fold cross-validation scheme. We then used TPE to randomly sample an initial uniform distribution of each of our hyperparameters and window sizes and evaluate the loss function for each random set of inputs. TPE then replaces this initial distribution with a new distribution built on the results from regions of the sampled distribution that minimize *L(x)*:

$$p(x|y) = \begin{cases} l(x) & if\ y < y^* \\ g(x) & if\ y \geq y^* \end{cases}$$

where y^* is a quantile γ of the observed y values so that $p(y < y^*) = \gamma$. Importantly, y^* is guaranteed to be greater than the minimum observed loss, so that some points are used to build *l(x)*. TPE then chooses candidate hyperparameters to sample as those representing the greatest expected improvement, *EI*, according to the expression:

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} y p(y) dy}{\gamma l(x) + (1-\gamma) g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1}$$

In order to maximize *EI*, the algorithm picks points x with high probability under $l(x)$ and low probability under $g(x)$. Each iteration of the algorithm returns x^*, the next set of hyperparameters to sample, with the greatest *EI* based on previously sampled points.

C.5 Training the classifier

The ECLAIR classifier was trained in three stages, using a custom wrapper of the scikit-learn (Pedregosa et al., 2011) random forest (Breiman, 2001) classifier to allow for use of TPE to search both algorithm hyperparameters and residue window sizes simultaneously. In all cross-validations performed, we allowed TPE to search the following hyperparameters, beginning with uniform distributions of the indicated ranges: (1) minimum samples per leaf (0-1000), (2) maximum fraction of features per tree (0-1), and (3) split criterion (entropy or gini diversity index). The number of estimators (decision trees) in each random forest was fixed at either 200 for training the feature selection classifiers, or 500 for training the full ensemble. We also allowed TPE to search over residue window sizes (± 0-5 residues for a total window of up to 11 residues, centered on the residue of interest). This was achieved by allowing extra features for neighboring residues to be included at the time of algorithm initialization.

In the first stage of training, cross-validation using TPE was performed on classifiers trained using only features from 1 of the 5 feature categories. The feature or set of features from each category with the minimum loss was selected to represent that category in building the ensemble classifier. In the second stage, the ensemble classifier was built of 8 random forest classifiers, each trained on different subsets of feature categories, and hyperparameters and window sizes were again chosen using cross-validation and TPE. In the final stage, following performance measurement on the testing set, the 8 sub-classifiers were retrained using the full set of 3,447

interactions with at least 50% UniProt residue coverage in the PDB, using the same hyperparameters and window sizes found in the previous step.

C.6 Evaluating the ensemble

After training and optimizing using only the training set, we predicted interface residues in the testing set. For each sub-classifier of the ensemble, all residues in the testing set that could be predicted (given the full set of necessary features or a superset of necessary) were ranked according to their raw prediction scores and ROC and precision-recall curves produced.

C.7 Benchmarking against other methods

Interface predictions for proteins in our testing set were accessed for several popular methods (de Vries and Bonvin, 2011; Kufareva et al., 2007; Liang et al., 2006; Porollo and Meller, 2007) through the CPORT web portal (de Vries and Bonvin, 2011). We compiled a set of representative protein structures from the PDB for each protein in our testing set, selecting the structure with the highest UniProt residue content based on SIFTS and excluding any PDB structures of interacting protein pairs from our testing set. We then evaluated the precision, recall, and false positive rate for proteins that were able to be classified by all methods. These represent point estimates of these metrics for the external methods with binary prediction scores.

C.8 Predicting new interfaces

We retrained the ensemble using the hyperparameters found during training and using all available co-crystal structures, including those from both testing and training sets. Using this fully trained ensemble of classifiers, we predicted interface residues for the remaining 118,113 interactions not

resolved by either PDB structures or homology models. Sub-classifiers were ordered based on the number and information content of each classifier. Each residue was then predicted by only the top ranking classifier of the ensemble trained on the full set or a subset of available features for that residue. To improve the separation of scores, we applied a consistent linear scaling operation to all raw prediction values whereby all scores up to the ~99.9th percentile of all prediction scores were rescaled from 0 to 1 and all scores above were capped at 1. We then evenly divided the predictions into 5 interface potential categories: Very Low (0.0-0.2), Low (0.2-0.4), Medium (0.4-0.6), High (0.6-0.8), and Very High (0.8-1.0).

C.9 Disease mutation analysis

We accessed all missense mutations from the Human Gene Mutation Database (HGMD) (Stenson et al., 2014) to compute the log odds ratio:

$$LOR = ln\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right)$$

where p_1 is the probability of a disease mutation being at the interface and p_2 is the probability of any residue being at the interface. We computed the log odds ratio for residues in each of the interface prediction potential categories. We also computed the log odds ratio for interactions with known interfaces from PDB co-crystal structures, defined as all known interface residues from NACCESS calculations and all residues in Pfam (Punta et al., 2012) domains with ≥ 5 interface residues.

C.10 Mutagenesis validation experiments

We performed mutagenesis experiments in which we introduced random population variants from the Exome Sequencing Project (Fu et al., 2013) into known and predicted interfaces. Mutations were introduced into human proteins according to our previously published Clone-seq pipeline (Wei et al., 2014) and their impact was assessed using our yeast two-hybrid assay. We tested the impact of 2,164 mutations on 1,167 interactions and computed the fraction of interactions disrupted in both known interfaces and predicted interfaces in the Low – Very High categories.

C.11 Web server

ECLAIR is deployed as an interactive web server (http://eclair.yulab.org) containing known and predicted interfaces for 130,634 protein interactions in 8 species. For each interaction, the most reliable, high-resolution model is presented, i.e. co-crystal structures are always displayed in lieu of homology models, and all remaining unresolved interactions are predicted by our classifier. Co-crystal structures are derived from the PDB, with extraneous chains removed for each interaction, and homology models are computed by modeler and downloaded from Interactome3D (Mosca et al., 2013). For both types of structural model, we computed all residues at the interface over all available models, and allow users to view any model from which a unique interface residue has been calculated. For predicted interfaces, a non-redundant set of individual protein models are shown when available. In total, the resource contains 7,135 interactions with co-crystal structures, 5,386 with homology models, and 118,113 with predicted interfaces.

C.12 REFERENCES

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research *25*, 3389-3402.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E.*, et al.* (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res *40*, W597-603.

Bergstra, J.S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. Paper presented at: Advances in Neural Information Processing Systems.

Berman, H.M. (2000). The Protein Data Bank. Nucleic Acids Research *28*.

Breiman, L. (2001). Random Forests. Mach Learn *45*, 5-32.

Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875-1882.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

de Vries, S.J., and Bonvin, A.M. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PloS one *6*, e17695.

Fu, W., O'Connor, T., Jun, G., Kang, H., Abecasis, G., Leal, S., Gabriel, S., Rieder, M., Altshuler, D., Shendure, J.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007). PIER: protein interface recognition for structural proteomics. Proteins *67*, 400-417.

Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J Mol Biol *55*, 379-400.

Liang, S., Zhang, C., Liu, S., and Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. Nucleic Acids Res *34*, 3698-3707.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295-299.

Morcos, F., Hwa, T., Onuchic, J.N., and Weigt, M. (2014). Direct coupling analysis for protein contact prediction. Methods Mol Biol *1137*, 55-70.

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nature methods *10*, 47-53.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.*, et al.* (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825-2830.

Pieper, U., Webb, B., Barkan, D., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E., Pettersen, E., Huang, C.*, et al.* (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research *39*, 74.

Pierce, B.G., Hourai, Y., and Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PloS one *6*, e24657.

Porollo, A., and Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. Proteins *66*, 630-645.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7*, 539.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet *133*, 1-9.

UniProt-Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res *43*, D204-212.

Velankar, S., Dana, J., Jacobsen, J., van Ginkel, G., Gane, P., Luo, J., Oldfield, T., O'Donovan, C., Martin, M.-J., and Kleywegt, G. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research *41*, 9.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.